

DataNet Full Proposal: DataNetONE (Observation Network for Earth)

I. Personnel.

PI: William Michener, U. New Mexico (UNM)

Co-PIs: Robert Cook, Oak Ridge National Laboratory (ORNL); Mike Frame, U.S. Geological Survey (USGS) National Biological Information Infrastructure (NBII); Stephanie Hampton, National Center for Ecological Analysis and Synthesis (NCEAS), UC-Santa Barbara; Kathleen Smith, National Evolutionary Synthesis Center (NESCent), Duke U.

Co-Is (Core Cyberinfrastructure Team): Paul Allen, Cornell; Jeffery Horsburgh, Utah State U.; Matthew Jones, NCEAS; Robert Sandusky, U. Illinois-Chicago; Ryan Scherle, NESCent; Mark Servilla, UNM; Dave Viegla, U. Kansas; Bruce Wilson, ORNL

Co-Is (Working Group and Education/Outreach Leaders and International Participants): Suzie Allard, U. Tennessee; Peter Buneman, U. Edinburgh; Randy Butler, UIUC-NCSA; John Cobb, ORNL; Patricia Cruse, California Digital Library (CDL); Ewa Deelman, USC-ISI; David DeRoure, U. Southampton; Cliff Duke, Ecological Society of America; Carole Goble, U. Manchester; Donald Hobern, CSIRO, Australia; Peter Honeyman, U. Michigan; Vivian Hutchison, NBII; Steve Kelling, Cornell U.; Jeremy Kranowitz, The Keystone Center; John Kunze, CDL; Bertram Ludaescher, UC Davis; Lorraine Normore, U. Tennessee; Ricardo Pereira, Brazil; Line Pouchard, ORNL; Carol Tenopir, U. Tennessee; Jake Weltzin, USGS; Von Welch, UIUC-NCSA

II. Intellectual Merit. To address the growing environmental, social, and technological challenges facing the world, scientists, educators, librarians, resource managers, and the public need open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

DataNetONE is designed to provide the distributed framework, sound management, and reliable technologies which enable the long-term preservation of diverse and complex multi-scale, multi-discipline, and multi-national science data. DataNetONE will initially emphasize multi-disciplinary observational data collected by biological (genome to ecosystem) and environmental (atmospheric, ecological, hydrological, and oceanographic) scientists, national and international research networks, and environmental observatories. However, the DataNetONE structure is not domain-specific, and will be extended to serve a broader range of science domains both directly and through interoperability with other DataNet deployments. Long-term preservation and access to data are assured by the DataNetONE approach of building a sustainable distributed network that leverages the considerable, multi-decade expertise of participating organizations (i.e., archives, libraries, environmental observing systems and research networks, science synthesis centers, and professional societies) and that develops and adopts agile, robust technologies and interoperability solutions. This stable infrastructure will facilitate the development and adoption of standards-based tools enabling users to discover, visualize, and integrate archived and current data. Economic and technical sustainability will be enabled by the DataNetONE External Advisory Committee and the DataNetONE International User Group that includes key mission agencies, businesses, research enterprises, professional societies, libraries, and individuals.

III. Broader Impacts. DataNetONE will accomplish its mission by making the scientist an active member of the data preservation process. Working Groups, led by the Co-PIs and Co-Is, engage a broad and diverse group of graduate students, educators, and leading computer, information, and library scientists, and will: (1) perform cutting edge computer science, informatics, and social research; (2) develop DataNetONE interfaces and prototypes; (3) adopt/adapt interoperability standards; (4) create value-added technologies (e.g., semantic mediation, scientific workflow, and visualization) that facilitate data integration, analysis, and understanding; (5) address sociocultural barriers to sustainable data preservation and data sharing; and (6) promote the adoption of best practices for managing the full data life cycle. Education and outreach are integral. DataNetONE engages graduate students in research and cyberinfrastructure development, and supports diverse participation by US and international collaborators. Outreach efforts will: (1) provide training to scientists and students via outreach to professional societies and through “Best Data Practices” presentations that will annually reach thousands of scientists who participate in NBII, NCEAS, NESCent, LTER, ORNL, and DataNetONE activities; (2) educate and engage citizen scientists in the full data life cycle through their involvement in the USA National Phenology Network and numerous Cornell Laboratory of Ornithology citizen science efforts; and (3) provide web-based multi-media training in managing, preserving, analyzing and visualizing Earth observational data.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.C.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) (Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)	15	
References Cited	3	
Biographical Sketches (Not to exceed 2 pages each)	64	
Budget (Plus up to 3 pages of budget justification)	85	
Current and Pending Support	46	
Facilities, Equipment and Other Resources	5	
Special Information/Supplementary Documentation	0	
Appendix (List below.) (Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

DataNet Full Proposal: DataNetONE (Observation Network for Earth)

Preface. Nothing less than a new type of distributed organization is required to provide easy and perpetual access to the data needed by scientists, decision-makers, and citizens to understand the nature and pace of change on Earth and to address associated environmental, social, and technological challenges. DataNetONE encompasses a diverse, enthusiastic, and capable international team of scientists sharing a common cyberinfrastructure (CI) vision that supports the preservation, sharing, and use of Earth observational data for decades to come. DataNetONE represents our collective view of how such a new organization can provide the distributed framework, sound management, coordinated and focused research, and robust technologies that will enable sustainable, long-term preservation of diverse multi-scale, multi-discipline, and multi-national Earth observational data. The DataNetONE enterprise is designed from its inception to be scalable (i.e., size of holdings, domains represented, countries served), flexible (i.e., engaging and learning from a broad spectrum of academic, library, research, business, and citizen-based organizations, including other DataNet partners) and sustainable (i.e., building on the long-term CI experience of its member organizations and adopting the most promising economic and technological sustainability solutions).

Roadmap to this Proposal. **Section 1** provides examples of prior NSF research, CI development, and education efforts by the five Co-PIs that are most directly relevant to DataNet (additional projects supported by NSF and other agencies for all 35 Co-Is are described in *Appendix A6*). **Section 2** introduces the DataNetONE rationale and vision, and highlights the economic and technological sustainability strategies that will be explored and implemented as part of our long-term Sustainability Plans, which are further documented in *Appendix A1*. **Section 3** summarizes the distributed DataNetONE architecture, which includes member nodes, coordinating nodes, and investigator toolkits, and describes how the enterprise will accommodate multiple evolving metadata and interoperability standards (details about the CI and its capabilities are included in *Appendix A3*). **Section 4** illustrates the DataNetONE organizational, governance, and Working Group structure, which are more fully described in *Appendix A2*, and lists many of the initial national and international collaborators whose specific contributions to DataNetONE are summarized in *Appendix A5* and *Letters of Collaboration*. **Section 5** describes how members of the Core CI Team and Working Groups will perform research, engage the community, and create and sustain DataNetONE (individual Co-I roles and qualifications are detailed in *Appendix A4* and the enclosed CVs). **Section 6** highlights the project's numerous intellectual merits and broader impacts.

1. Results of Prior NSF Support for DataNetONE Principal Investigators

DataNetONE Co-Principal Investigators (**BOLD**) and Co-Investigators (underlined) have a rich history of collaboration in information technology research, providing informatics services to the scientific community, and educating an informatics-literate work force.

Cyberinfrastructure and Information Technology Research: **W. Michener**, M. Jones, B. Ludaescher, D. Viegla, and others from 8 institutions located in the US and Scotland [NSF 0225665, "Enabling the Science Environment for Ecological Knowledge (SEEK)", \$12,250,000, 10/1/02 – 9/30/08] created CI for ecological, environmental, and biodiversity research, and provided ecoinformatics education to junior faculty. New CI included an integrated data grid (EarthGrid) for accessing ecological and biodiversity data and analytical tools, as well as Kepler, an open-source scientific workflow solution; see <http://seek.ecoinformatics.org/> and <http://kepler-project.org/>. **M. Frame** and individuals from three other Institutions [NSF ITR-0326460 "Science on the Semantic Web – Prototypes in Bioinformatics", \$2,396,001, 9/1/03 – 8/31/08] are developing a framework to facilitate science research and education on the semantic web, and implementing and evaluating prototype tools and applications for use in the biocomplexity and biodiversity domains. The testbed for prototyping capabilities is the NBII web portal (<http://www.nbii.gov/>).

Scientific Synthesis and Informatics Support: The Network Office of the US LTER (LNO; NSF DEB-0236154, \$9,011,235, 3/1/03 – 2/28/09; R. Waide, J. Brunt, **W. Michener**, J. Vande Castle) provides administrative, CI, training, and scientific synthesis support for the network of 26 Long Term Ecological Research sites located in the US, Puerto Rico, French Polynesia, and Antarctica (<http://www.lternet.edu/>). LNO (M. Servilla, **W. Michener**, and others) is developing a Network Information System to support the next decade of synthetic science across the network and that will include an array of diverse sites in addition to LTER; the prototype effort is focused on EcoTrends (see <http://trends.sagehost.net/index.php>). The National Center for Ecological Analysis and Synthesis (NCEAS; NSF 0553768, \$18,402,599, 10/1/06

– 9/30/11, O.J. Reichman , **S. Hampton**) uses collaborative synthesis and analysis to address issues of importance to ecology and environmental biology and to disseminate this information to researchers, policy-makers, and the broader community. The award funds Working Groups, Distributed Graduate Seminars, sabbatical faculty, and postdoctoral fellows at NCEAS and provides partial funding for informatics staff (M. Jones and others) who have pioneered new CI for managing ecological data, including Ecological Metadata Language (EML) and the Knowledge Network for Biocomplexity (KNB), a data sharing and replication network among interlinked research sites. The National Evolutionary Synthesis Center (NESCent; NSF EF-0423641, \$15,000,000, 12/1/04 – 11/30/09; **K. Smith**) supports administrative, outreach, and informatics staff (R. Scherle, and others), over 20 resident postdoctoral and sabbatical scholars, and approximately 700 scientists participating in working groups and other meetings sponsored by the Center (<http://www.nescent.org/index.php>). NESCent's informatics group leads CI initiatives in open source database, software, and semantic web technologies for evolutionary biology. The education and outreach program translates the results of evolutionary biology to the education community and general public and helps recruit evolutionary biologists from underrepresented groups.

Education, Outreach, and Community-Building: In addition to outreach and training efforts that they support through their affiliated institutions (e.g., LTER, USGS-NBII, NCEAS, NESCent, ORNL-DAAC), many of which are focused on reaching under-represented groups, project investigators have participated in research coordination networks (NSF 0639794, 0129792, 006567) that have coalesced research communities, pioneered new informatics sub-disciplines, provided informatics training, and developed community databases. **R. Cook**, **M. Frame**, **J. Weltzin**, **B. Wilson**, and others participate in the “USA National Phenology Network RCN” [USA-NPN, NSF 0639794; M. Schwartz and 9 others] creating a nationwide network of phenological observations. RCN activities are defining the CI (i.e., database and information systems) to enable researchers and interested citizens to store, discover, and retrieve phenological data from this distributed database network. The USA-NPN actively supports student/early researcher exchange/training programs, as well as participation by students, citizens, and scientists from under-represented communities in all network activities. **W. Michener** led a “Research Coordination Network: Resource Discovery Initiative for Field Stations (RDIFS)” [NSF DBI-0129792, \$500,000, 2001-2006] whereby the LTER Network Office, Organization of Biological Field Stations (OBFS), NCEAS, and numerous other institutions collaborated to facilitate storage, discovery, and access to the strategic environmental information resources that are collectively held at North American biological field stations. Activities included implementing the OBFS Data Registry (which now contains more than 4,200 items), developing several key OBFS databases, revamping the OBFS web site (<http://www.obfs.org/>), and infusing informatics and geospatial technologies into field stations and marine laboratories. This latter activity was enabled via a series of innovative two-week-long training sessions that were held annually to disseminate software tools, knowledge, and resources. More than 140 field station personnel from more than a third of the field stations representing the United States, Canada, Costa Rica, Panama, the Bahamas, and French Polynesia participated in the training. **R. Cook** participates in “FLUXNET: A Global Network Measuring and Assessing Spatial-Temporal Variability of Carbon Dioxide, Water Vapor and Energy Fluxes between the Terrestrial Biosphere and Atmosphere” [(NSF 006567; D. Baldocchi and 9 others], which provides data and knowledge on surface fluxes that are needed by scientists from atmospheric chemistry, biogeography, ecohydrology, ecosystem ecology and biogeochemistry. Eddy flux measurements and the meteorological, soil and plant variables measured at each site produce information that will drive, parameterize, and validate the next generation of models that predict ecosystem balances of carbon, water and nutrients, and weather and climate. The FLUXNET Data System provides flux data files in standardized formats (comma delimited text or netCDF) using standard parameter names and units. The project serves the international community by convening international workshops and hosting visiting scientists. FLUXNET is instrumental in training the next generation of biogeoscientists from developed and developing countries by hosting a young scientist forum.

2. Introduction: Rationale, Vision, and Approaches to Achieving Sustainability

The rationale for DataNetONE is simple: *People of all countries are experiencing increasing environmental, social, and technological challenges associated with climate variability, altered land use, population shifts, and changes in resource availability (e.g., food, water, and oil). Scientists, educators, librarians, resource managers, and the public need open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. Such data are critical as they form the basis for good scientific decisions, wise management and use of resources, and informed decision-making.* Here,

we summarize how DataNetONE is envisioned to meet this need and describe many of the economic and technological approaches that will be employed to ensure that the organization can realize its vision for decades into the future.

2.1 DataNetONE Vision. DataNetONE provides the distributed framework (which is comprised of Member and Coordinating Nodes as illustrated in Figure 1), sound management, and robust technologies that enable long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data. DataNetONE initially emphasizes observational data collected by biological (genome to ecosystem) and environmental (atmospheric, ecological, hydrological, and oceanographic) scientists, research networks, and environmental observatories. DataNetONE will be domain agnostic, progressively expanding to broader domains and building on infrastructure and interoperability with DataNet partners.

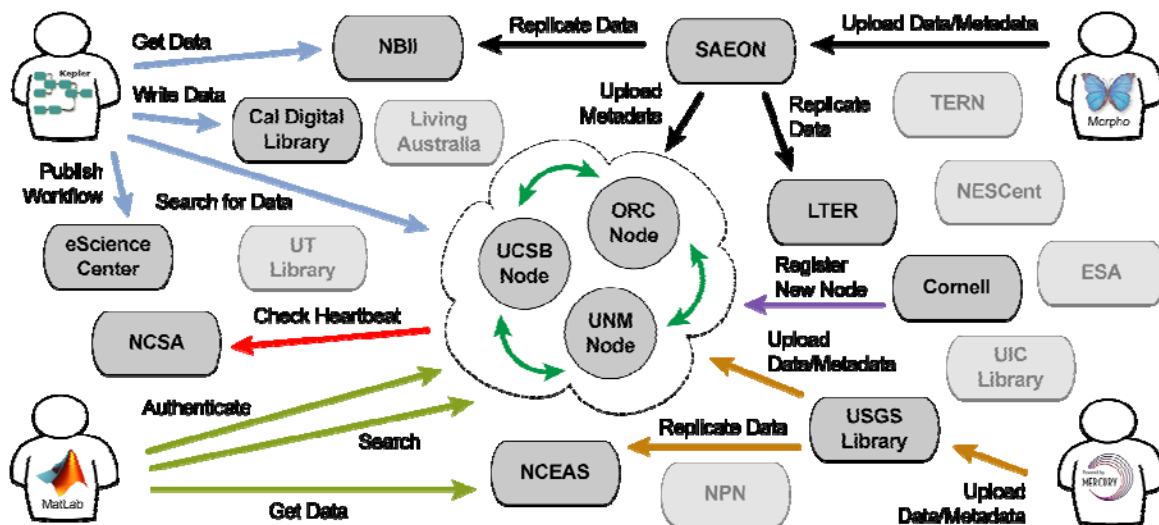


Figure 1: DataNetONE Member Nodes form a robust, distributed network via coordinating services provided by a set of Coordinating Nodes (i.e., Oak Ridge Campus, UC-Santa Barbara, and University of New Mexico) arranged in a high-availability configuration. Scientists and citizens interact with Member Nodes (e.g., South African Environmental Observation Network, California Digital Library, USGS National Biological Information Infrastructure) through software tools that utilize standardized interfaces. This structure supports many different usage scenarios, such as data and metadata management and replication (e.g., using Morpho [black arrows] or the Mercury system [orange arrows]), as well as analysis and modeling (e.g., using commercial software like Matlab [light green arrows] and open-source scientific workflow systems like Kepler [blue arrows]). Coordinating Nodes perform many basic indexing and data replication services to ensure data availability and preservation (e.g., node registration [purple arrow] and monitoring via heartbeat services (red arrow)).

DataNetONE is not the end, but rather the means to enable scientists and citizens to address and better understand the difficult and complex biological, environmental, social, and technological challenges affecting human, ecosystem, and planetary sustainability. The comprehensive cyberinfrastructure allows novel questions to be asked that require harnessing the enormity of existing data and developing new methods to combine and analyze diverse data resources (Figure 2).

DataNetONE will accomplish its goals by making scientists, students, librarians, and citizens active participants in the data life cycle—especially, the data preservation process. By supporting community-derived interoperability standards and incorporating new value-added and innovative technologies (e.g., for semantic and geospatial information, scientific workflows, and advanced visualization) into the scientific process, DataNetONE will facilitate sophisticated data integration, analysis, interpretation, and understanding. A strong education and outreach program focuses on scientists and students learning to better and more easily manage, preserve, analyze, and visualize Earth observational data. Citizen scientists are actively engaged in data preservation and scientific discovery through their involvement in programs such as the USA National Phenology Network (USA-NPN) and numerous Cornell Laboratory of Ornithology citizen science efforts (e.g., eBird, Project FeederWatch).

DataNetONE Services

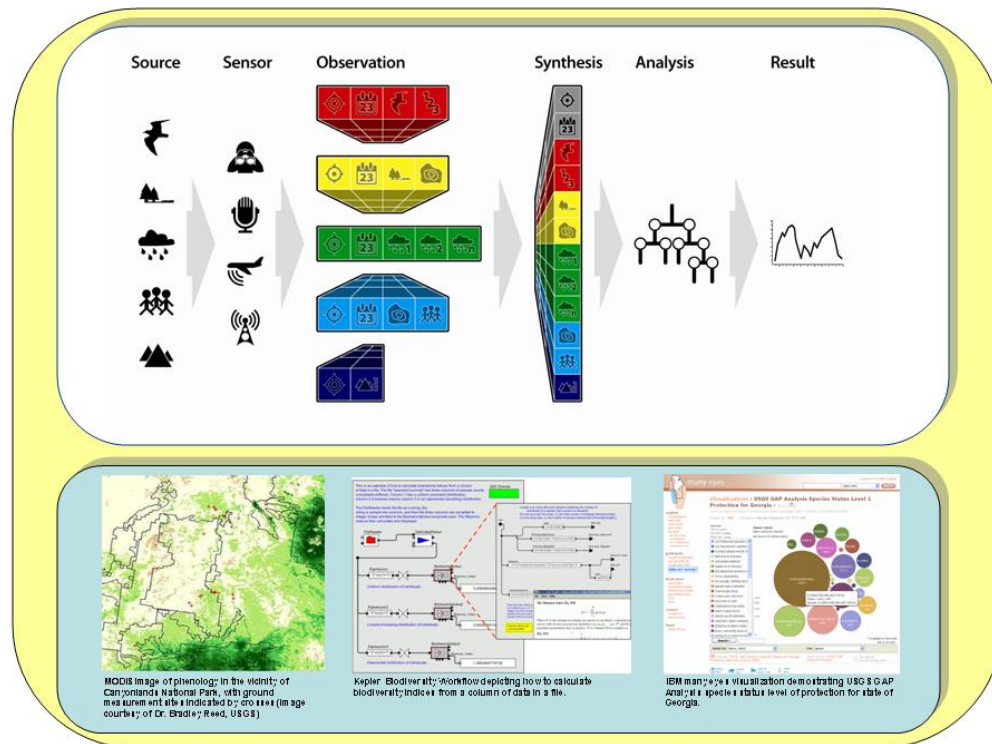


Figure 2: DataNetONE value-added services enable scientists to address novel questions through integrating disparate data sources (top), supporting geospatial data processing (lower left), and providing scientific workflow solutions like Kepler (lower middle) and high-level analyses and visualizations like IBM ManyEyes.

2.2 Strategies for Achieving Long-term Economic and Technological Sustainability. Long-term preservation and access to data are assured by the DataNetONE approach – a sustainable distributed network that leverages the considerable commitment, expertise, and funding support of participating organizations (i.e., archives, libraries, environmental observing systems and research networks, science synthesis centers, government agencies, and professional societies) and that develops and adopts agile, robust technologies and interoperability solutions. For example, Coordinating Nodes are associated with:

- **University of New Mexico (UNM).** The Long-Term Ecological Research Network Office at UNM (LNO) coordinates activities among a network of 26 research sites extending from Alaska to Antarctica and has been funded since 2000 by NSF, USDA-ARS, and the USDA-FS. The New Mexico Computing Applications Center (NMCAC) is supported by the State of New Mexico and Intel Corporation. Intel Corporation is hosting the DataNetONE infrastructure which will be situated on the National LambdaRail and Internet2 at Intel's Rio Rancho facility (24/7 support) and co-located with NMCAC's SGI/Intel High Performance Computer system (172 teraflops). UNM will provide secure office and conference space at the UNM Science & Technology Park.
- **Oak Ridge Campus (ORC).** The University of Tennessee (UT) partnered with Oak Ridge National Laboratory (ORNL) to establish the Joint Institute for Computational Sciences (JICS) to advance scientific discovery and state-of-the-art computer science. JICS is housed in a new 52,000 sq. ft. building, constructed by the state of Tennessee on ORC. The JICS team includes UT-ORNL Joint Faculty appointees (e.g., Bruce Wilson, DataNetONE Co-I), JICS Research Affiliates, postdoctoral fellows, graduate students, and administrative staff. Computing resources for DataNetONE will be purchased through UT and housed on the Oak Ridge Campus. JICS provides an innovative and productive organizational structure to support advanced research and CI for the nation.
- **University of California Santa Barbara (UCSB).** The National Center for Ecological Analysis and Synthesis (NCEAS) at UCSB has stimulated synthetic and collaborative research in ecology and environmental science since its inception in 1995. NCEAS has facilitated research on over 400

synthesis projects involving over 4,000 researchers from many science disciplines. NCEAS ranks in the top 1% of 38,000 institutions publishing in ecology and the environment in terms of scientific citation [1]. The highly collaborative NCEAS Ecoinformatics group is complemented by an Ecoinformatics cluster at UCSB, involving data managers and programmers from NCEAS, LTER and other large data-intensive ecological research programs. The intersection of these vibrant groups of resident Ecoinformatics researchers and thousands of resident and visiting scientists in ecology and allied disciplines increasingly present special opportunity for synergy at NCEAS.

Likewise, Member Nodes are initially to be located on six continents in association with an array of organizations that have been supported through governmental and non-governmental sources (Figure 3).

In addition to sustained funding associated with the Coordinating and Member Nodes, DataNetONE's sustainability strategy explores a varied portfolio of income streams including: (1) tiered dues-paying membership in the organization; (2) a project-driven model, with an array of corporate, foundation, and government agencies supporting specific projects; (3) an endowment; (4) a fee-for-service model where new DataNetONE or existing applications are developed or customized for a data user's specific needs; (5) conference fees; (6) administrative fees from new grants; and (7) a Wikipedia model with contributors offering time and resources (see Appendix A1 for more detail on these seven options). DataNetONE will also seek input from venture capital and organizational sustainability experts participating in the DataNetONE Sustainability Working Group (cf. Sections 4 and 5).

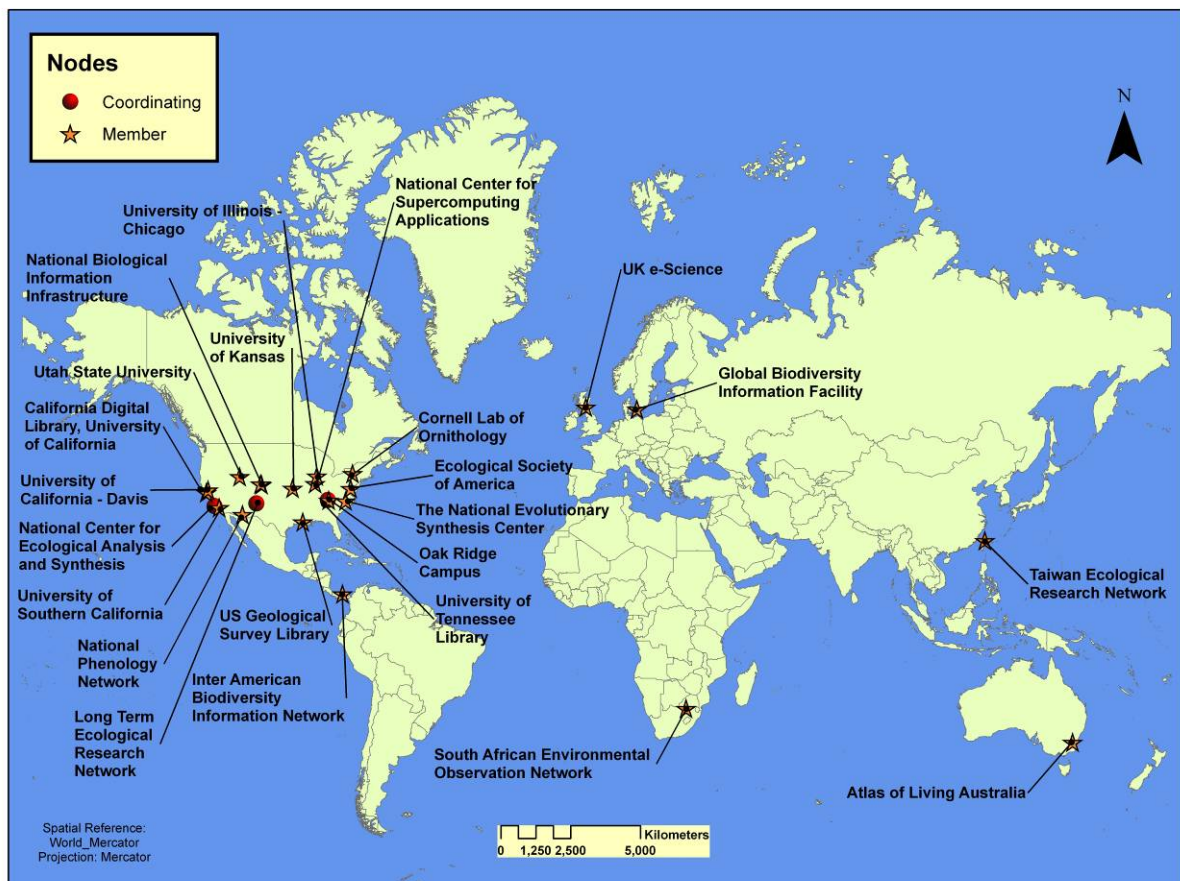


Figure 3: Geographic distribution of initial Coordinating and potential Member Nodes of DataNetONE. Coordinating Nodes are geographically dispersed and orchestrate replication of data from originating Member Nodes to geographically dispersed nodes.

To promote technological sustainability and manage systems migrations, DataNetONE is based on a distributed systems architecture, with redundant service providers, metadata repositories, and data storage nodes. The Core Cyberinfrastructure Team (CCIT) will monitor, evaluate, and test evolving

systems, technologies, and standards in development nodes. DataNetONE will also use commercial service providers where appropriate, which will further enable transparent technology migrations.

3. Overview of DataNetONE Architecture. The DataNetONE architecture must embrace the highly dispersed and independent nature of data collection activities relevant to the environmental and earth sciences. Data are collected by tens of thousands of scientists around the world who have the expertise to describe and archive these data, as well as curate them. Attempting to centralize this curation function is inherently untenable and will not scale. Thus, DataNetONE will achieve both scalability and sustainability through a highly distributed system architecture (Figures 1 and 4) that utilizes the **DataNet Service Interface** to access uniform services provided and used by three types of cyberinfrastructure: (1) **Member Nodes** located at institutions distributed throughout academia, libraries, government agencies, and other organizations that provide local data storage, curation, and metadata for a set of data resources that are collected or affiliated with that institution; (2) **Coordinating Nodes** that are geographically-distributed (Figure 3) to provide a high-availability, fault-tolerant, and scalable set of coordinating services to the Member Nodes, including a complete metadata index and data replication services for all data in all Member Nodes; and (3) an **Investigator Toolkit** that provides a complete and evolving set of tools for data and metadata management by scientists and curators throughout the entire data life cycle (Figure 3). Initially, there will be three Coordinating Nodes geographically dispersed at ORC, UNM, and UCSB. A small number of additional coordinating nodes may be implemented as DataNetONE expands in scope, sustainable funding, and international presence.

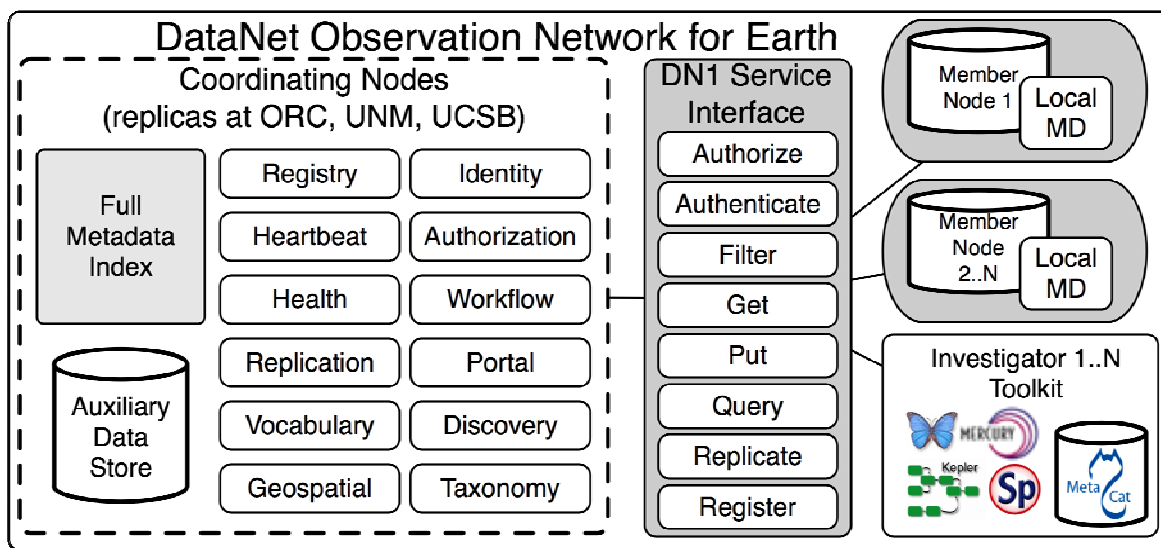


Figure 4: Main features of the DataNetONE architecture, emphasizing distributed data storage at Member Nodes and metadata indexing and services provided by Coordinating Nodes. Communication between Member Nodes (e.g., for replication), between Member Nodes and the Investigator Toolkit (e.g., for inserting data), and between Member Nodes and Coordinating Nodes (e.g., for metadata indexing) is all mediated via a common DataNetONE Service Interface that spans all node types.

All DataNetONE systems will communicate using the shared **DataNetONE Service Interface**, which defines services for the identification, authentication, and authorization of participants, registration and tracking of Member Nodes, and data management operations such as data archiving, search, and data integration. Communication using the DataNetONE Service Interface will be frequent between the Member Nodes and Coordinating Nodes, but it will also be common between pairs of Member Nodes (e.g., in order to stage data for researchers at different locations) and between the Investigator Toolkit and Member nodes (e.g., to document and archive data). The DataNetONE Service Interface will enable groups to choose their level of participation in DataNetONE based on their capabilities and the services they are able to provide to users. A set of fundamental services for communication with the Coordinating Nodes, along with a defined service level agreement, will be needed for full member node status. The

DataNetONE Service Interface will include Network Services, Federated Identity and Authorization Services, Object Management Services, Preservation Services, and Discovery and Usage Services.

The Network Services are implemented at Coordinating Nodes to track the existence and state of Member Nodes. The services include a *Registry Service* that tracks available Member Nodes, a *Heartbeat Service* to monitor liveness of the nodes, a *Health Service* for determining the current state of nodes in the network, and a *Capacity Service* for monitoring available storage and network resources throughout DataNetONE. Formal systems monitoring, analysis, and visualization systems will be deployed that extend current network and systems monitoring frameworks that are already in use in the computing centers of the Coordinating Nodes.

The Federated Identity and Authorization Services will be created to protect the privacy and confidentiality of contributed data. These services include a federated *Identity Provider Service* that spans the broad community of data producers and consumers in DataNetONE, *Authentication Services*, and *Authorization Services* to enforce community governed "data use" policies. This system will ensure that policies regarding access to data are consistently enforced regardless of their physical location. At the highest level, DataNetONE must define standards and mechanisms for communicating "data use" policies set forth by contributors, as well as address policies for data embargoes on time-scales that extend beyond the life of the contributor. These policies address not only who can access data, but requirements for logging that data access. To enforce the data use policies, the DataNetONE infrastructure must support authorization for data resources in a distributed environment where local control of resources will persist. DataNetONE must also seamlessly support the movement of individuals between different organizations, and the corresponding changes in authorization that may accompany such changes. To allow for risk mitigation, since many other aspects of the DataNetONE infrastructure rely on security, the Federated Identity and Authorization Services will be staged. Initially, simple identity-based access control lists will be used; this will be enhanced over time with finer-grained policies. Initial enhancements will include role-based access control and policies with regard to logging.

Many authentication systems exist today, including "push" sequence (Kerberos [2,3], Keynote [4]), "pull" sequence (Radius [5], RSVP [6]), "agent" sequence (GSI [7]), and hybrid systems that use combinations of the above [8]. DataNetONE will pursue models developed as part of the Grid Security Infrastructure [7] and incorporate community or role-based authorization where necessary to resolve scale issues of many users to many resources, which can quickly degrade performance. Approaches used in Metacat [9] and other network systems (e.g., LDAP directory services (X.500), X.509 certificates, or the Shibboleth [10] identity system) may provide an initial starting point, but the Federated Identity Working Group (Section 5.2.1) must research solutions similar to those used by Grid community (e.g., MyProxy [11,12], GridShib [13]) that support identity management across a broad landscape of users and resources. Federated identity management is an ongoing area of active research and there are currently no ideal alternatives although both SAML [14] and OpenID [15] appear to provide enticing solutions with various implementations available (e.g., Shibboleth). In both cases it will still be necessary to provide adapters for the data client tools to authenticate against the DataNetONE identity provider, and to provide a replicated implementation of the identity provider.

A particular challenge is integration of the various identity solutions already in use by researchers with the DataNetONE Identity Provider. Our approach is to leverage existing sources of user identity to the greatest extent possible; e.g., we will leverage the InCommon federation and Shibboleth technologies. For the portions of our use community that fall outside of that realm, we will leverage third party identity providers (e.g., [16]) and run our own identity server, as a last resort. Using Shibboleth, allows us to also leverage existing group memberships where compliant with our authorization policies.

The Object Management Services are used to transfer data among nodes and from the various client tools. These services will support basic CRUD operations to Create, Read, Update, and Delete objects. These services will conform to existing protocols wherever possible to maximize utility of existing client tools (e.g., GridFTP, OPeNDAP).

The Preservation Services will be used by the Coordinating Nodes to guarantee long-term preservation and viability of DataNetONE resources. The *Replication Service* will actively track replicas for digital objects, ensuring that sufficient geographically redundant copies exist on viable Member Nodes. The *Migration Services* will continuously inspect archived objects to identify those associated with data and metadata standards approaching end-of-life and migrate these objects to current standards. The *Validation Service* will help verify conformance to community-accepted standards for data and metadata. Finally, the *Digital Identifier Service* will be used to create and register naming authorities for assigning

accession numbers for objects and managing versioning using standards such as LSID [17] and ARK [18].

The Discovery and Usage Services support searching, use, and interpretation of data located at Member Nodes. The Discovery Service will support standard query languages (e.g., XQuery, SPARQL, etc.) used in existing metadata catalogs and deploy advanced semantic search systems developed by the Data Integration and Semantics Working Group (Section 5.2.5). The *Logging and Notification Service* and the *Provenance Service* provide detailed, policy-driven logging of access and data lineage that produces a coherent view of data usage and derivation regardless of which node serviced the data requests. The *Workflow Service* provides mechanisms for creating, archiving, sharing, and communicating about workflows created in systems such as Kepler and mechanisms for executing these workflows that will consume and produce data in DataNetONE. The *Ontology Service* provides mechanisms for creating shared vocabularies, taxonomies, and geospatial gazetteers for use in metadata standards and the semantic annotation of digital objects.

3.1 Member Nodes. As part of the sustainability of DataNetONE, the storage of data sets will be distributed across Member Nodes. One copy will typically reside at the originating Member Node, and replicas will be created at two or more other locations, such as other Member Nodes, the Coordinating Nodes, commercial providers such as Amazon S3, and the rapidly evolving world of cloud storage such as the planned Google Science Storage capability. Member Nodes will include the broad array of science stakeholders, including University libraries, research networks like the Long Term Ecological Research Network and the Organization of Biological Field Stations, synthesis centers like NCEAS and NESCent, government agencies like the USGS and NASA, and emerging environmental observatories like NEON, WATERS, and OOI. Many institutions have a substantial investment in existing data management infrastructure, so the requirements for participating as a Member Node will be specified as a set of Service Interfaces that must be implemented at each Member Node. This allows the nodes to either utilize their existing infrastructure by providing the required Service Interface implementations, or they can deploy the Member Node software stack provided by DataNetONE to create these services. Thus, each node may provide a different subset of the services depending on the needs of their clients and their existing infrastructure. Also, each Member Node will be scaled according to the needs of its client base, and will contribute differing levels of resources to the DataNetONE collaboration. Smaller nodes such as an individual field research station may only provide modest storage resources, but DataNetONE gains significantly in the aggregate from many such nodes. Larger initiatives such as observatories and agencies will bring substantial resources and will benefit from the coordination with the rest of the science community that DataNetONE provides. DataNetONE envisions twelve or more Member Nodes throughout the world by year 3 of the project and anticipates accelerated growth thereafter.

3.2 Coordinating Nodes. Coordinating Nodes provide critical network-wide services to the DataNetONE Member Nodes and to users of DataNetONE. The foremost of these services are: 1) a registration service to track which Member Nodes are participating and are online; 2) a global metadata index spanning all data resources held by Member Nodes; 3) an identity provider service and authorization service for establishing trust and security relationships among Member Nodes and their users; 4) a replication service that tracks and maintains multiple replicas of all data objects across available Member Nodes; and 5) a discovery service for locating data resources of interest to scientists (see Appendix A3).

Earth observation data sets available today are relatively modest in size. For example, 92% of data sets at the ORNL DAAC are less than 100MB in size. The larger data sets arise mainly from satellite imagery and sensor networks. To initially bootstrap the distributed DataNetONE network, the Coordinating Nodes will also serve as Member Nodes and provide data storage services using storage arrays acquired as part of the initial hardware purchases. We expect the size and number of data sets to grow significantly as sensor networks and observatories such as NEON and WATERS come online. However, as the number of Member Nodes grows, we expect they will assume a larger responsibility for data storage, while the relative role of the Coordinating Nodes for data storage will decrease.

The software stack for coordinating nodes will implement the DataNetONE Service Interface using virtualization of both servers and storage. This software stack will be developed in conjunction with Member Nodes and users as open-source software modules that can be deployed in a range of DataNetONE nodes, including the Coordinating Nodes, the data center Member Nodes, and even into individual researcher laboratories. Implementation of this software will be accelerated by adopting and adapting tools already in use within DataNetONE participating organizations, such as the NCEAS and

LTER Metacat data server [9, 19], the ORNL and NBI Mercury toolset for metadata editing, indexing and searching [20], and the Kepler workflow engine [21]. We will build on current work in the library and information science communities, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [22], and in the scientific community on data exchange standards, such as the SEEK EarthGrid [23] and OPeNDAP [22] protocols. By providing the coordination, oversight, and resources to fuse these tools into a coherent operational platform, DataNetONE will both provide the value back to these member organizations to justify their participation in DataNetONE, as well as catalyze the formation of an interoperable and sustainable framework for data management.

3.3 Investigator Toolkit. Active participation of the individual scientific researchers is critical to the success of DataNetONE. The Investigator Toolkit provides a desktop and web-based software stack that makes participation in the network simple for researchers and small institutions. For individual researchers, the stack will include a group-oriented data management package designed by the CCIT and the Sociocultural Issues Working Group that allows researchers, their students, and colleagues to easily document and share the products of their scientific activities with their immediate local colleagues. This software system will allow them to easily synchronize their daily activities with a Member Node of their choosing, such as their local university library. Researchers will use the toolkit to assign metadata about data sets, to describe and archive their analytical procedures and scientific workflows, and to integrate DataNetONE data into their analytical and modeling approaches. The toolkit will adopt and adapt existing tools such as the Morpho metadata editor [24], the Specify system for managing collections data, and the Kepler system for scientific workflows [21], as well as provide interfaces to common analytical tools such as Matlab and R. The researchers gain significant local advantage due to the increased organization and accessibility of their own data for their own research purposes while simultaneously enabling new collaborations through data sharing with their colleagues and science communities, and preserving their data for future generations of science.

3.4 Multiple Evolving Standards. In earth science disciplines, thousands of individual scientists manage billions of heterogeneous data sets and observations, contribute to hundreds of community specific database islands with idiosyncratic schemas and data formats, and utilize several dozen metadata specifications (e.g., EML [25], CSDGM [26], BDP [27], ISO19115 [28], GML [29]) [30]. There is no one-size-fits-all model and no single site institution, or even virtual organization that can manage this diversity centrally. For example, relevant data sets span small-scale ASCII representations of tabular relational data, meso-scale spatial and temporal time series from various *in situ* instruments represented in self-describing formats like HDF, NetCDF, and OPeNDAP, and large-scale remote sensing data from satellites and other instruments represented as raster-grids. Metadata specifications that are widely used overlap tremendously in their purpose and coverage (e.g., Ecological Metadata Language (EML) [25], Earth Science Markup Language (ESML), and the FGDC Biological Data Profile). In addition, these metadata specifications continue to evolve themselves, introducing new versions regularly. We do not expect this situation to improve substantially over the short term, and so DataNetONE will be designed to accommodate multiple, evolving metadata standards and data representation standards as determined by the user communities served. Some consistency will be achieved by mapping the most common discovery metadata dealing with theme, coverage (geospatial, temporal, and taxonomic), and attribution as represented in all of the standards. By utilizing schema-independent metadata middleware like Metacat, we expect DataNetONE to be able to simultaneously address the temporal evolution of specifications, diversity of community-accepted standards, and long-term viability of the DataNetONE holdings. The latter will be accomplished through active management of the lifecycle of specifications such that each specification used in the community will progress through periods of testing, active deployment, deprecation, and obsolescence. As specifications move towards obsolescence, automated mappings to migrate existing resources to current standards will be developed. Because this is never a lossless process, original metadata and data records will be maintained but superseded by new versions of the data and metadata resources. The evolution of data and metadata standards is such a critical issue to long-term persistence of data that these topics will be the foci of research working groups within DataNetONE, including the Data Preservation, Metadata, and Interoperability Working Group (Section 5.2.3) and the Data Integration and Semantics Working Group (Section 5.2.5).

Together, these elements form a robust distributed data network. The combination of centralized coordination with distributed participants will enable flexible, staged adoption of new technologies and standards, anticipating and responding to rapid technological and sociological change. The architecture is

designed to support the full data life cycle by providing a comprehensive set of data management tools for the researcher and enabling seamless migration of data from an investigator to appropriate DataNetONE nodes. This distributed, service-based model is designed to scale, enabling national and even global data management for diverse sciences both independently and through integration with other DataNet partners.

DataNetONE will implement this architecture and component tools in collaboration with current and future data providers. DataNetONE will help transform these providers into a coherent and sustainable data network while enabling discipline-specific interfaces and support. Development and integration of an appropriate toolkit for the data generators and consumers will bring both groups fully into the data preservation and access life cycle. While this architecture addresses the technological issues in data preservation, DataNetONE recognizes that a complete solution to barriers affecting data preservation must also address organizational and sociocultural issues, as discussed in the following sections.

4. DataNetONE Structure and Composition

4.1 Organizational Structure. DataNetONE is a multi-tiered organization (Figure 5) and will be governed by an External Advisory Committee (EAC) chartered to ensure that the organization is fulfilling its mission, serving its user communities, implementing a sustainable business plan, and developing an evolving, strategic vision for the enterprise. EAC members will have a range of expertise (Technical, Management / Corporate, Data Provider / Data User, Professional Society, Library) and be recognized as international leaders in these areas. DataNetONE will be directed and supervised by the Executive Director (ED) who is responsible for all technical, management and budget issues. The ED will report to the EAC and UNM Vice President for Research, interact with the NSF Program Director(s) for DataNet, and coordinate with other DataNet partners. The ED will also evaluate and implement strategies for economic sustainability.

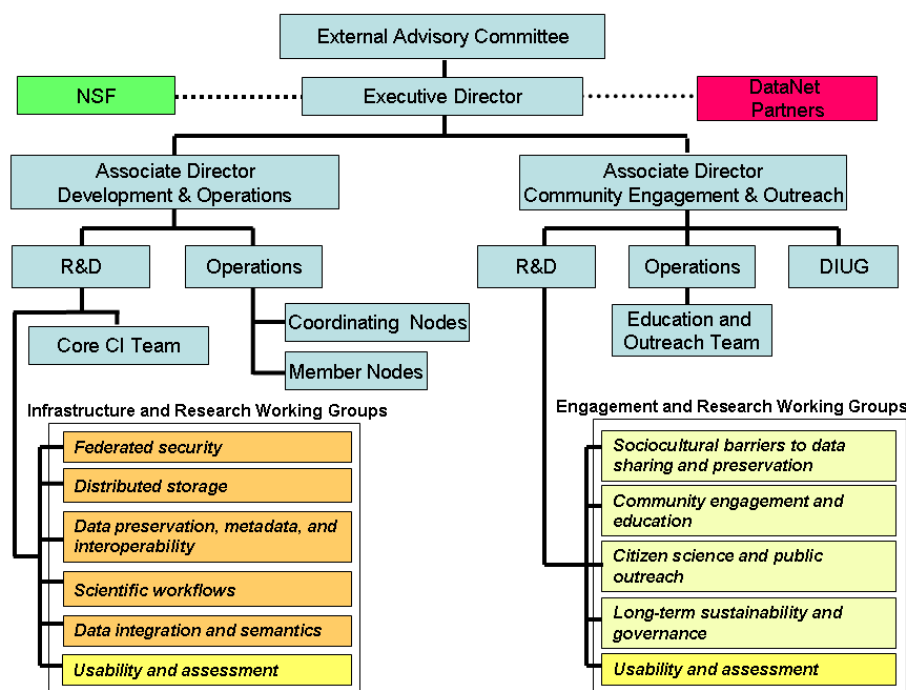


Figure 5: DataNetONE organizational structure. Bold lines depict management hierarchy within DataNetONE and dashed lines show ties with NSF and other DataNet projects. The DataNetONE International User Group (DIUG) is the worldwide community of Earth observation data authors, users, and diverse stakeholders.

The Associate Director for Development and Operations (AD D&O) will oversee development and implementation of architecture, computer science research, and technological evolution through the activities of Working Groups and the CCIT (Section 5), including the staff of full-time developers. The CCIT will be co-directed by B. Wilson (ORC) and M. Jones (UCSB), and initially includes D. Vieglais (U

Kansas), J. Horsburgh (Utah State, CUAHSI), R. Scherle (NESCent), R. Sandusky (UI-Chicago), P. Allen (Cornell), and M. Servilla (UNM). The AD D&O also provides a crucial role in facilitating communication and knowledge transfer between concurrent and temporally disjoint Working Groups [31].

An Associate Director for Community Engagement and Outreach (AD CE&O) will oversee activities of the Office and staff, and be responsible for subcontracting and procurement. The AD CE&O will engage the community through the DataNetONE International User Group (DIUG), which consists of Earth observation data authors, users, and diverse stakeholders including students, educators, researchers, libraries, data centers, professional societies, policy-makers, and the general public. DIUG will meet annually to identify the evolving technical challenges and opportunities that can be applied to advance education, research, and policy through the use of DataNetONE data products, tools, and services. The DIUG will advise DataNetONE of needs and suggestions from its members (i.e., “requests”). Requests identified by DIUG will be prioritized by the AD CE&O and AD D&O and tasked to the CCIT or relevant Working Group(s) (Sect. 5). A request tracking system will keep the DIUG informed about progress. Other web-based mechanisms support communication and engagement throughout the year.

A comprehensive Assessment and Evaluation Program, including formative and summative evaluation, will be initiated by the Usability and Assessment Working Group (Section 5.2.6) at the start of the project to ensure that desired products are delivered on time and that broad community involvement occurs throughout the project. Work packages based on a logic modeling approach [32] will be established for all Working Groups and the CCIT to allocate resources, establish realistic milestones, track the success of all supported activities, and follow accepted project management guidelines [33].

4.2 Composition. DataNetONE participants represent a wide range of sectors from across the nation and globe. The **domain** focus is on the Earth observation sciences including the breadth of relevant biological (genomics, biodiversity, and ecological) and environmental (atmospheric, hydrological, and oceanographic) sciences. Key partners and affiliations include: **academic institutions** from the US (including three EPSCoR states—Tennessee, Kansas, and New Mexico) and the United Kingdom (i.e., Edinburgh, Manchester, Southampton); **research networks** (e.g., LTER, CUAHSI, Taiwan Ecological Research Network, South African Environmental Research Network (SAEON)); environmental observatories (e.g., NEON, USA-NPN, Ocean Observatory Initiative, South African Environmental Observatory Network); **NSF- and government-funded synthesis** (i.e., NCEAS, NESCent, Atlas of Living Australia) and **supercomputer centers/networks** (ORNL, NCSA, and TeraGrid); **governmental organizations** (e.g., USGS, NASA, EPA); **academic libraries** (e.g., University of California Digital Library, University of Tennessee, and University of Illinois-Chicago libraries, which are active in the digital library community and are members of the Coalition for Networked Information, the Digital Library Federation, and the Association of Research Libraries); **international organizations** (e.g., Global Biodiversity Information Facility, Inter American Biodiversity Information Network, Biodiversity Information Standards); numerous large **data and metadata archives** (e.g., USGS-National Biological Information Infrastructure, ORNL Distributed Active Archive Center for Biogeochemical Dynamics, World Data Center for Biodiversity and Ecology, Knowledge Network for Biocomplexity); **professional societies** (e.g., Ecological Society of America, Natural Science Collections Alliance); **NGOs** (e.g., The Keystone Center); and the **commercial sector** (e.g., Amazon, Battelle Ventures, IBM, Intel).

In addition to the above participants, DataNetONE’s users include: (1) scientists and students throughout the world who will be provided with tools for data authoring and analysis; (2) librarians, library educators, and undergraduate and graduate educators who will use DataNetONE in their own research as well as for educating the next generation of scientists and science librarians; and (3) K-12 educators, students and citizens, whose activities may range from writing simple reports to developing educational programs and active participation in citizen science programs. DataNetONE will engage these users in the organization activities (Working Groups and DIUG) and by creating dialogue through education and outreach activities. DataNetONE policies and practices will promote broad inclusion of all stakeholders in the DIUG and will strive to increase the participation of women and individuals from groups underrepresented in science and engineering.

5. DataNetONE Activities

The CCIT and a series of focused but evolving Working Groups will perform research, create and maintain the CI, engage the broad stakeholder community, and contribute to developing a sustainable governance and funding plan.

5.1 CCIT Activities. The CCIT: (1) designs and directs the implementation of DataNetONE systems with participating organizations and other DataNet Partners, linking local data providers into a seamless network that supports unified protocols and standards for identifying, describing, contributing, searching, and accessing metadata and data; and (2) collaborates with the research Working Groups. The CCIT will persist indefinitely within DataNetONE to support the evolving technology landscape and will operate under several fundamental design principles. First, DataNetONE will be founded on open standards and protocols for interoperability, and will produce open source products. Second, partnering organizations will develop key toolset components and DataNetONE will actively engage other open source resources, such as the Google Summer of Code. Third, DataNetONE will build on current work in the library and information science communities on system interoperability and data exchange standards (e.g., Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH, 22], OPeNDAP [34]) and data curation and preservation [35-39].

DataNetONE activities will address the complete data life cycle through a comprehensive program of research, design, and development to create a system to *preserve*, *disseminate*, and *protect* research objects in a secure, reliable, and open system that is responsive to users' and scientists' needs.

5.1.1 Data Deposition, Acquisition, and Ingestion. DataNetONE will preserve data by addressing the deposition, acquisition, and ingestion of data and associated descriptive (domain), administrative (system), structural, and preservation metadata. Investigators have traditionally resisted depositing data in centralized repositories, so DataNetONE must ensure that data acquisition integrates directly into the day-to-day work of scientists through continual community engagement that ensures that tools and systems are relevant to scientists and other community members. Providing value to individual scientists and participating organizations is critical to ensuring the contributions of those organizations to the vision and mission of DataNetONE. We plan to adapt existing data and metadata management tools that are used in various disciplines (e.g., Morpho [24] and eBird [40]) to produce a standards-based data management toolkit that can be deployed in scientific labs. This evolving investigator's toolkit will provide a mechanism for organizing investigator data locally (e.g., among graduate students in a lab) while enabling seamless contribution to the DataNetONE distributed repository. We anticipate this toolkit will include both web-based and application-based tools for scientists to enter data and associated metadata and the capability to archive raw data streams or rich media from automated sensor networks for backup or sharing purposes. A key aspect of our approach is the standards and protocols upon which this toolkit is based, which will enable users, partners, and the general community to augment the tools which DataNetONE provides, so that the toolkit can include both the specialty tools needed by various scientific disciplines and the more general tools that are discipline agnostic. Deposited data artifacts must be uniquely identified and versioned [41,42] at an economically feasible scale using emerging standards such as Life Science Identifiers [LSID, 17, 43], Archival Resource Keys [18], or Digital Object Identifiers [DOI, 44] that emphasize location-independent identification. Transparent policies, procedures, and best practices for scientists, data authors, and information managers will be developed from community input. DataNetONE will accomplish this by soliciting input from the stakeholder community, and through Working Groups. DataNetONE will engage domain scientists and students in the data and metadata ingest and creation process by providing training through academic and research libraries' education programs, synthesis centers and research networks, professional society events, and seminars that will annually reach thousands of scientists through DataNetONE partner activities.

5.1.2 Data Curation, Metadata Management. DataNetONE views data, analysis products, and metadata records as specializations of digital objects, and their curation and management present common challenges such as the heterogeneity challenges described in Section 3. DataNetONE's curation strategy is to support massively parallel local data management in the community by providing a shared framework for coordination and replication of data across distributed data providers and clients. We will adopt and adapt existing coordination interfaces (e.g., the Open Geospatial Consortium's map services [WFS, WCS, WMS, 45], and the EarthGrid service interfaces [23]) that have been used to provide a common communications interface between diverse data systems. These distributed data providers will be connected through a geographically redundant set of Coordinating Nodes that maintain a complete metadata index that can be used to rapidly discover, access, and manage data objects from the distributed providers (Figure 3). Coordinating Nodes will index data objects and annotations from local data providers (e.g., Metacat [9], Mercury/NBII [20], DiGIR/TDWG [46], and OPeNDAP [34]). By supporting multiple local management systems, we ensure that DataNetONE is flexible enough to deal

with technology evolution. Quality assurance procedures will be performed throughout the data life cycle, including continuous detection and correction of object corruption, and object replication using the LOCKSS [47] approach to ensure that multiple viable and geographically distributed copies of each object exist. Fine-grained object usage will be captured in order to support management, evaluation, and auditing. Because standards for object formats and metadata change over time, we will provide policies and tools for format migration, forward migration between versions of standards, adoption of new standards, and de-commissioning of retired standards. Curation requires close coordination with other activities such as data protection and metadata interoperability.

5.1.3 Data Protection. To protect the privacy and confidentiality of contributed data we will integrate authentication and authorization tools with community supported "data use" policies. First, DataNetONE must support a federated identity management system that spans the broad community of data producers and consumers. To decide whether an individual or organization is allowed to access data, identity of the individual must be established, which is problematic for a distributed system across organizations. DataNetONE will need to incorporate community or role-based authorization where necessary to resolve scale issues of many users to many resources, which can quickly degrade performance. Finally, DataNetONE will define standards and mechanisms for communicating "data use" policies set forth by contributors, and address policies for data embargoes on time-scales that extend beyond the life of the contributor. The Federated Security Working Group (Section 5.2.1) will research and propose solutions to these issues that can be implemented within DataNetONE.

5.1.4 Data Discovery, Access, Use, and Dissemination. DataNetONE will disseminate data by developing and encouraging third-party development of multiple data access tools that meet the needs of diverse communities. DataNetONE portals will provide the general ability to access, explore, and visualize DataNetONE's rich data resources using effective geospatial, temporal, taxonomic, and thematic search and browse features. More targeted tools and interfaces will allow DataNetONE products to be used within, for example, analysis and modeling applications (e.g., Kepler Workflow system, Matlab, R, etc.), simulation models, and citizen science gateways (e.g., eBird). We anticipate building social networking tools into the DataNetONE portal, aiding data discovery and facilitating collaboration. Success of DataNetONE will be measured not only by the volume of data organized, but also the number and variety of users, and frequency of user visits.

5.1.5 Data interoperability, standards, and integration. DataNetONE will: 1) promote the efficient use and continuing evolution of existing standards (e.g., metadata interoperability, ontologies, semantic frameworks, and knowledge representation strategies); 2) support community-based efforts to develop new standards and merge or adapt existing standards; and 3) provide systems, tools, procedures, and capacity to enhance data interoperability and integration. We will empower the community to self-organize and manage scientific data formats, database schemas, metadata frameworks, and ontologies through specific DataNetONE tools and procedures. We plan to model all schema-level objects as first-class knowledge artifacts, called "Digital Schema Objects" (DSO), similar to the "Content Model Objects" in the Fedora repository software [48]. We will design and implement the schema object repository architecture, research and develop schema classification and interoperability assessment services (e.g., BigDig [49], OBOE [50]), and research source registration, mapping, and integration approaches (e.g., using semantics [21,51], provenance [52], or bi-directional "lenses" [53]). The schema repository will require tools for both harvesting community DSOs (e.g., Darwin Core schemas via BigDig [42]) and accepting user-submitted DSOs. Data providers can search and adopt pre-existing schemas from the schema repository, customize existing schemas for the providers' needs, and contribute extensions back to the community.

5.1.6 Data Evaluation, Analysis, and Visualization. DataNetONE users will have access to diverse data resources without needing to understand the details of data storage or exchange formats. DataNetONE will support the discovery of available tools by including searchable tool metadata. We will support the creation of user-generated visualizations and analyses and provide a repository of user-generated objects that can be copied, modified, reused, or incorporated as components into other projects or presentations (e.g., myExperiment [54]). Many of these tools will be developed in collaboration with our partners and various commercial organizations, such as the IBM ManyEyes research project [55], the Kepler scientific workflow project [21], and others. Workflow systems [56] will be used to integrate disparate analytical systems, automate back-end data processing, coordinate grid systems, and design and implement

analysis interfaces for scientists. The tools will facilitate identifying distributions, patterns, and trends that advance scientific knowledge, develop effective public policy, improve the scientific and technical literacy of students, and enable more rapid analysis by various stakeholders.

5.1.7 Education and Training. Education and training will broadly increase understanding of the long-term value of data and the tools that aid data preservation, use, and re-use. Activities will be designed to seek community feedback on needs and usability. Working Groups will guide the development of products appropriate to general classes of users, and will develop and implement a strategy for dissemination. Rather than creating entirely new programs to address education and to seek user feedback, we will most effectively reach targeted communities through multiple existing citizen science, education, and outreach programs sustained by the organizations involved in DataNetONE (e.g., [57-60]).

5.1.8 Community and user input and assessment. The user community will be active participants in the shaping of policies, procedures, systems, and tools through direct integration in DataNetONE's management structure (Figure 4). DataNetONE will address this diverse group of users by conducting evaluative research to understand their current practices and future needs using ethnographic [61], human factors and usability analysis methods [62] as DataNetONE tools and products are deployed within Education and Training Activities. An annual evaluation plan will be produced with results, recommendations for products and services, and suggestions for future directions.

5.2 Working Groups on Research, Infrastructure, and Community Engagement. Working Groups are central to DataNetONE in conducting research, specifying CI, and engaging the community. The Working Group model will allow us to conduct targeted research and education activities with much broader groups of scientists and users than are directly involved in the DataNetONE proposal. Working Groups are also designed to enable research and education activities to evolve over time. As the topics for one Working Group are resolved, or as the needs of DataNetONE are better specified or change, the number and foci of the Working Groups will evolve. Each Working Group will have two co-leaders who are compensated to organize the activity and propose solutions to particular research, education, and cyberinfrastructure problems. Initial Working Group foci, including their leaders, selected participants, and a subset of anticipated activities and products are described below.

5.2.1 Federated security. Von Welch, Butler. i) establish federated identity management scheme; ii) establish authorization/access-control for provisioning resources within a distributed DataNetONE infrastructure that supports a large user-base; iii) adopt a standard licensing mechanism, such as "Creative Commons" for data and products within the DataNetONE environment.

5.2.2 Distributed storage. Cobb, Honeymann. i) define and select production-wide area file system(s); ii) define and select production data (file, block, storage object) movement services for transfer of data between nodes and for transfer to and from users; iii) define and select production data-related services including tools for file replication management, replication location, staging, and planning as well as the specification of needs for continuous validation, data warming, and consistency checking.

5.2.3 Data preservation, metadata, and interoperability. Kunze, Hobern, Pereira, Buneman. i) identify, evaluate, select, and implement the standards, tools, procedures, and internal policies needed to support data curation and preservation and metadata management; ii) exercise the standards, procedures, and tools deployed at the initial system implementation; iii) develop a plan for a comprehensive internal summative evaluation to determine the effectiveness of tools, procedures, and systems.

5.2.4 Scientific workflows. Goble, Deelman, De Roure. i) evaluate and co-develop workflow archival formats; ii) develop data and workflow provenance interoperability framework; iii) generalize existing, emerging workflow repositories; iv) gather/develop workflow design patterns for commonly used systems.

5.2.5 Data integration and semantics. Ludaescher, Pouchard. i) design schema object repository architecture; ii) specify schema classification and interoperability assessment services; iii) research and prototype source registration, mapping, integration services.

5.2.6 Usability and assessment. Frame, Normore. i) interact with DIUG Community and initiate research to assess current practices and future needs using user-centered design approach (e.g., initially survey users interacting with existing archives and metadata management systems in use at participating institutions for rapid input on usability issues); ii) recommend enhancements to tools, products, and services; iii) oversee assessment plan that assures deliverables and schedules are met, and that broad community involvement occurs throughout the project life cycle.

5.2.7 Sociocultural issues. *Allard, Tenopir*. i) identify and examine the sociological and cultural issues that inhibit effective data sharing and long-term preservation; ii) evaluate and recommend strategies that overcome sociocultural barriers and create incentives for data preservation; iii) explore and make recommendations regarding the roles for libraries in training data authors, supporting data curation, and acting as a facilitator of digital preservation practices.

5.2.8 Community engagement and education. *Hutchison, Hampton, Duke*. i) determine effective mechanisms for community input on tools for data providers and consumers and for the dissemination of products appropriate to scientific and non-scientific audiences; ii) establish a training program for both science and citizen science initiatives; and iii) establish metrics that will be used to determine the adoption success and utility of DataNetONE products.

5.2.9 Citizen science and public outreach. *Kelling, Weltzin*. i) determine requirements for management of citizen science data and visualization, exploration, and analysis of data by disparate users (from citizens to scientists); ii) create a comprehensive data management strategy for highly disparate citizen-based observational networks; iii) build tools to allow project managers, researchers, educators, or networks to develop a customizable web-based data gathering system.

5.2.10 Long-term sustainability and governance. *Kranowitz, Michener, Cruse*. i) investigate different organizational models, including a stand-alone non-profit 501(c)(3) organization; ii) investigate different funding models to ensure long term sustainability; iii) establish the governance of DataNetONE and a representative stand-alone organization (i.e., DIUG) to ensure that stakeholders provide direction.

6. Intellectual Merit and Broader Impacts

6.1 Intellectual Merit. DataNetONE will become the foundation for innovative new data intensive science through the creation of a stable distributed data organization structure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data from the biological and environmental sciences. Importantly, DataNetONE recognizes that proper data management is not an end but a means to provide proper access, synthesis, exploration, and analysis of the diversity of observational data. The organization leverages and expands upon computer science, informatics, and social research, as well as CI associated with leading international academic institutions, archives and data centers, libraries and digital libraries, NGOs, and governmental organizations. The organizational structure (i.e., DIUG and the Working Group model) ensures that DataNetONE is broadly representative of its stakeholders, sustainable, and adaptable to future needs. The comprehensive information infrastructure being developed will allow novel questions to be asked that require harnessing the enormity of existing data and developing new methods to combine and analyze diverse data resources. For example, understanding and moderating the increasing anthropogenic pressures exerted on ecological systems (e.g., global warming, habitat destruction, infectious disease transmission) will require a sufficiently robust information infrastructure to provide access to a diversity of data resources along with tools for data manipulation, exploration, analysis, and visualization.

6.2 Broader Impacts. Technology often plays a vital role in the emergence of new science. DataNetONE will organize observational data into a massive distributed repository, which provides the foundation for new data-intensive computing that will transform how the biological, ecological, and environmental sciences are carried out. DataNetONE will create a new research paradigm in which data are synthesized from diverse and wide ranging disciplines to provide insight into the patterns and trends from gene expression to entire ecosystem events. DataNetONE recognizes that digital data are not only the output of research, but the foundation for new scientific insights, and allows scientists, students, and citizens to become active participants in gathering, managing, accessing, exploring, and visualizing its rich data resources. Diverse participation by US and international collaborators and graduate students, as well as the inclusion of women and members of under-represented groups in DIUG and throughout DataNetONE will facilitate expansion and adoption of DataNetONE throughout the world. A strong web-based education program coupled with the numerous public outreach and citizen science programs supported by DataNetONE partners guarantees that thousands of scientists, students, and citizens will learn how to better manage, preserve, analyze and visualize Earth observational data.

References Cited

- [1] ISI Essential Science Indicators. 2005.
- [2] Kohl, J. and C. Neuman. 1993. The Kerberos network authentication service (V5). Internet RFC 1510.
- [3] Neuman, B. and T. Ts'o. 1994. Kerberos: An authentication service for computer networks. *IEEE Communications* 32:33–38.
- [4] Blaze, M., J. Feigenbaum, and A. Keromytis. 1999. KeyNote: Trust management for public-key infrastructures. *Proc. 1998 Cambridge Security Protocols International Workshop, LNCS*, 1550:59-63.
- [5] Rigney, C., S. Willens, A. Rubens, and W. Simpson. 1997. Remote authentication dial in user services (RADIUS), IETF RFC 2138.
- [6] Baker, F., B. Lindell, and M. Talwar. 2000. RFC 2747 - RSVP Cryptographic authentication, Standards Track RFC.
- [7] Butler, R., V. Welch, D. Engert, I. Foster, S. Tuecke, J. Volmer, and C. Kesselman. 2000. A national-scale authentication infrastructure. *Computer* 33:60-66.
- [8] Lorch, M., B. Cowles R. Baker, L. Gommans, P. Madsen A. McNab, L. Ramakrishnan, K. Sankar, D. Skow, and M. Thompson. 2004. Conceptual grid authorization framework and Classification, Global Grid Forum, <http://www.ggf.org/documents/GFD.38.pdf>
- [9] Berkley, C., M. Jones, J. Bojilova, and D. Higgins. 2001. Metacat: a schema-independent XML database system. *Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM '01)*, Fairfax, VA, pp. 171-179.
- [10] Erdos, M. and Cantor, S. 2002, "Shibboleth Architecture DRAFT v05. <http://shibboleth.internet2.edu/docs/draft-internet2-shibboleth-arch-v05.pdf>, May 2002.
- [11] Novotny, J., S. Tuecke, and V. Welch. 2001. An online credential repository for the Grid: MyProxy. *HPDC-10*, p. 104
- [12] Basney, J., M Humphrey, and V. Welch. 2005. The MyProxy online credential repository. *Software: Practices and Experience* 35: 801-816.
- [13] Barton, T., J. Basney, T. Freeman, T. Scavo, F. Siebenlist, V. Welch, R. Ananthakrishnan, B. Baker, M. Goode and K. Keahey. 2006 Identity federation and attribute-based authorization through the Globus Toolkit, Shibboleth, GridShib, and MyProxy", 5th Annual PKI R&D Workshop, 14 p.
- [14] SAML. <http://www.opensaml.org/>
- [15] OpenID. <http://openid.net/>
- [16] www.protectnetwork.com
- [17] LSID (Life Sciences Identifier): <http://lsids.sourceforge.net/>
- [18] ARK: Archival Resource Key <http://www.cdlib.org/inside/diglib/ark>
- [19] Jones, M.B., C. Berkley, J. Bojilova, and M. Schildhauer. 2001. Managing scientific metadata. *IEEE Internet Computing* 5: 59-68.
- [20] Mercury: <http://mercury.ornl.gov/>
- [21] Ludäscher B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao. 2006. Scientific workflow management and the Kepler System. *Special Issue: Workflow in Grid Systems. Concurrency and Computation: Practice & Experience* 18:1039-1065.
- [22] OAI-PMH: <http://www.openarchives.org/>
- [23] Michener, W.K., J.H. Beach, M.B. Jones, B. Ludaescher, D.D. Pennington, R.S. Pereira, A. Rajasekar, and M. Schildhauer. 2007. A knowledge environment for the biodiversity and ecological sciences. *Journal of Intelligent Information Systems* 29:111-126.

- [24] Higgins, D., C. Berkley, and M.B. Jones. 2002. Managing heterogeneous ecological data using Morpho. *Proceedings of the 14th International Conference on Scientific and Statistical Database Management, July 24-26, 2002*. J. Kennedy (ed). ISBN 0-7695-1632-7 ISSN 1099-3371.
- [25] Fegraus, E.H., S. Andelman M.B. Jones, and M. Schildhauer. 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Amer.* 86:158-168. [also see EML Ecological Metadata Language: <http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html>]
- [26] FGDC CSDGM Federal Geospatial Data Committee Content Standard for Digital Geospatial Metadata: <http://www.fgdc.gov/metadata/csdgm/>
- [27] Frondorf, A., M.B. Jones, and S. Stitt. 1999. Linking the FGDC geospatial metadata content standard to the biological/ecological sciences. *Proceedings of the Third IEEE Computer Society Metadata Conference*. Bethesda, MD. April 6-7, 1999.
- [28] ISO 19115 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020
- [29] Cox, S. et al. (eds.). 2003. *Open Geospatial Geography Markup Language (GML) Implementation Specification, Version 3.00*, Open Geospatial Consortium document 02-023r4 [Online: <http://www.opengis.org/docs/02-023r4.pdf>, 6 July 2004].
- [30] Jones, M.B., M.P. Schildhauer, O.J. Reichman, and S. Bowers. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annu. Rev. Ecol. Evol. Syst.* 37:519–544.
- [31] Reich, B.H. 2007. Managing knowledge and learning in IR projects: a conceptual framework and guidelines for practice. *Project Management Journal* 38(2):5 - 17
- [32] Kaplan, S.A. and K.E. Garrett. 2005. The use of logic models by community-based initiatives. *Evaluation and Program Planning* 28:167-172.
- [33] IEEE Std 1490™-2003, *Adoption of PMI Standard: A Guide to the Project Management Body of Knowledge*.
- [34] Cornillon, P., J. Gallagher, and T. Sgouros. 2003. OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal* 2:164-174.
- [35] Borghoff, U.M., P. Rödiger, J. Scheffczyk, and L. Schmitz. 2006. *Long-term Preservation of Digital Documents: Principles and Practices*. Berlin: Springer.
- [36] Friedlander, A., and P. Adler. 2006. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*. Association of Research Libraries.
- [37] Jantz, R. and M.J. Giarlo. 2005. Digital preservation: Architecture and technology for trusted digital repositories. *D-Lib Magazine*, 11(6) <http://www.dlib.org/dlib/june05/jantz/06jantz.html>
- [38] Lougee, W., et al. 2007. *Agenda for Developing E-science in Research Libraries*. Association of Research Libraries.
- [39] Wendler, R. 2006. The status of preservation metadata in the digital library community. In M. Deegan, & S. Tanner (Eds.), *Digital preservation* (pp. 60-77). London: Facet Publishing.
- [40] Kelling, S., B. Sullivan, and C. Wood 2007. What is eBird. <http://ebird.org/content/ebird/news/whitepaper.html>
- [41] Clark, T., S. Martin, and T. Liefeld. 2004. Globally distributed object identification for biological knowledgebases. *Brief in Bioinformatics* 5:59.
- [42] Hilse H-W. and Kothe, J. 2006.. *Implementing persistent identifiers : overview of concepts, guidelines and recommendations*. London: Consortium of European Research Libraries; Amsterdam : European Commission on Preservation and Access, 2006. 57 s. Dostupný také z WWW (URN): <<http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8>>. ISBN 90-6984-508-3.
- [43] Pereira, R., D. Hobern, R. Hyam, L. Belbin, K. Richards, and S. Blum. TDWG Life Sciences Identifiers Applicability Statement. <http://www.tdwg.org/standards/150/>

- [44] DOI: The Digital Object Identifier System <http://www.doi.org/hb.html>
- [45] WxS; see: WCS Web Coverage Service, <http://www.opengeospatial.org/standards/wcs>; WFS Web Feature Service, <http://www.opengeospatial.org/standards/wfs>; WMS Web Mapping Service, <http://www.opengeospatial.org/standards/wms>
- [46] TAPIR: TDWG Access Protocol for Information Retrieval, <http://wiki.tdwg.org/twiki/bin/view/TAPIR/>
- [47] Maniatis, P., M. Roussopoulos, T.J. Giuli, D.S.H. Rosenthal, and M. Baker. 2005. The LOCKSS peer-to-peer digital preservation system. *ACM Trans. Comput. Syst.* 23, 1 (2005), 2-50. <http://www.lockss.org/>
- [48] Fedora: www.fedoracommons.org
- [49] Vieglais, D. 2006. The Big Dig – A monitoring service for Distributed Generic Information Retrieval data services. <http://bigdig.ecoforge.net/>
- [50] Madin, J. S., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. 2007. An ontology for describing and synthesizing ecological observational data. *Ecological Informatics*, 2:279-296. doi:10.1016/j.ecoinf.2007.05.004
- [51] Pouchard, L.C., D.E. Bernholdt, and A. Woolfe. 2005. Data Grid Discovery and Semantic Web Technologies for the Earth Sciences. *International Journal on Digital Libraries* 5:72-83.
- [52] Green, T.J., Z.G. Ives, G. Karvounarkis, and V. Tannen. Update exchange with mappings and provenance, Intl. Conf. on Very Large Databases (VLDB), Vienna, Austria, September 2007.
- [53] Foster, J.N., M.B. Greenwald, C. Kirkegaard, B.C. Pierce, and A. Schmitt. Exploiting schemas in data synchronization. 2007. *Journal of Computer and System Sciences*.
- [54] myExperiment : <http://www.myexperiment.org/>
- [55] IBM ManyEyes: <http://services.alphaworks.ibm.com/manyeyes/home>
- [56] Taylor I., D. Gannon, E. Deelman, and M. Shields (eds.). 2007. *Workflows for eScience: Scientific Workflows for Grids*, Springer.
- [57] Andelman, S.J., C.M. Bowles, M.R. Willig, and R.B. Waide. 2004. Understanding environmental complexity through a distributed knowledge network. *BioScience* 54:240–246
- [58] Dolan, E.L., B.E. Soots, P.G. Lemaux, S.Y. Rheed, and L. Reiser. 2004. Strategies for avoiding reinventing the precollege education and outreach wheel. *Genetics* 166:1601-1609.
- [59] Paul, E. 2000. The National Biological Information Infrastructure: Present and future. *BioScience* 50: 22.
- [60] Powell, K. 2007. Breaking with tradition. *Nature* 446: 226-228.
- [61] Van House, N., M.H. Butler, and L.R. Schiff. 1998. Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. *Proceedings of CSCW '98: The ACM Conference on Computer Supported Cooperative Work*, Seattle, WA. 335-343.
- [62] Lazar, J. (ed.) 2007. *Universal Usability: Designing Computer Interfaces for Diverse User Populations*. Chichester; Hoboken, NJ: John Wiley & Sons.
- [63] Hutcheson, J.O. "The End of a 1,400-Year-Old Business," *Business Week*. Apr. 16, 2007. http://www.businessweek.com/smallbiz/content/apr2007/sb20070416_589621.htm
- [64] <http://www.w3.org/Consortium/>
- [65] http://www.iabin-us.org/projects/financial_sustainability/financial_sust_study.html
- [66] <http://hul.harvard.edu/jhove/>
- [67] Jungle Drive. <http://www.jungledisk.com/>
- [68] Apple Time Machine. <http://www.apple.com/macosx/features/timemachine.html>

University of New Mexico Central Office and Coordinating Node

DataNetONE Central Office at the UNM Science & Technology Park. The Central Office will be situated in secure space formerly occupied by the Joint Technology Office at 901 University SE (Figure 1). The space (3,330 ft²) includes seven offices, a large open space that can support several workstations/desks, a large conference room, rest rooms, and storage area. A lobby, kitchen, and walled-in patio are adjacent to the facility. Four additional conference rooms (available for Working Group meetings), a café, and ample parking are associated with the Park. High-speed internet access is provided by the University of New Mexico.



Figure 1: UNM Science & Technology Park (left) and DataNetONE Central Office floor plan (right).

Coordinating Node at New Mexico Computing Application Center at Intel Corporation Headquarters. The UNM DataNetONE Coordinating Node will be co-hosted with the New Mexico Computing Applications Center (NMCAC) at the Intel Corporation fabrication facility in Rio Rancho, New Mexico. Intel Corporation has operated their New Mexico facility since 1980 and opened the “Fab 11X” production unit in October 2002 as part of their 4 million feet² campus. The “Fab 11X” plant is Intel's first high-volume 300mm fully automated production operation facility that includes the Intel® Pentium® 4 processors and other advanced chips, and will be the production site for their next generation 45 nm processors. The NMCAC is home to the world's 3rd fastest supercomputer (2008), a 172 teraflops Altix® ICE system developed by Silicon Graphics, Inc. (SGI) and named “Encanto” (enchanted), with 14,336 Intel Xeon® processors, 28 Terabytes (TB) of memory, and 172 TB of online storage (Figure 2). The NMCAC system is supported as an integral part of Intel's Data Center infrastructure, including continuous operational support (conditioned power and environmental control) and fiber-optic connectivity to the National Lambda Rail network. The shared facility with NMCAC/Intel will optimize functional requirements of the DataNetONE node by taking immediate advantage of the redundant and fault-tolerant environment required by NMCAC/Intel.



Figure 2: The “Encanto” (enchanted) supercomputer located at the New Mexico Computing Application Center (NMCAC) and hosted by the Intel Corporation. facility in Rio Rancho, New Mexico.

Additional LTER Network Office Facilities to Support Conferences, Training, and Working Groups. The Long Term Ecological Research Network Office (LNO) occupies a 2,700 square-foot suite comprising seven offices, an 8-person technical workspace, and two 40-person conference rooms in the CERIA building on the main campus of University of New Mexico. For collaborative technology, the LNO supports a Polycom MGC50+ IP video conferencing bridge that can support video conferences of up to 48 persons. In addition, there are Polycom units that can be easily relocated in any of the working group conference facilities described above.

The LNO and the SEEK ITR project have co- dedicated a modern information technology training laboratory (Figure 3) that complements the above facilities. The training laboratory is optimized for student-to-instructor communication, while remaining ergonomically comfortable for long periods of instruction. The center piece of this laboratory is a fire-wall protected, 24-student pod facility with the latest Dell duo-core

Pentium desktop computers for each student, including dual 20 inch flat-screen monitors that can be shared through the instructor's computer and multimedia/video system. Administration of the training laboratory is performed by a full-time system and desktop support analyst.



Figure 3: LNO Informatics Training Lab.

The LTER Network Office hosts computer facilities for the LTER Network Information System Infrastructure; the backbone of which is the Network Office Data Center. This climate-controlled center has scalable servers and enhanced network bandwidth to better serve the LTER Network and its partners in the ecological community. 8 Dell Quad-Core PowerEdge servers with over 12 Terabytes of disk storage, redundant power supplies and UPS) serve as the core communication, collaboration, and data processing, storage, and delivery components of the LTERnet.edu domain. In addition, there are modern multi-processor development and test machines. The combination of Linux and Windows operating systems on the Intel platform allows for maximum flexibility in

incorporating new developments and technology. The Center standardizes on both PostgreSQL and MySQL relational database management systems, although Microsoft SQL Server is available for special purposes. In addition, the Center has a number of large format color output devices and a variety of scanning data input devices.

The UNM campus is wired with a 10 Gigabit redundant fiber backbone for optimal intra-campus networking needs. The CERIA building, which houses the LNO, have both fiber and copper Gigabit ethernet networking capability. Research activities at UNM enjoy OC-3 fiber connection to the Internet II via Denver that is connected directly to national Gigabit backbone infrastructures. In addition, the UNM is a full member of the National LambdaRail consortium. National LambdaRail a major initiative of U.S. research universities and private sector technology companies to provide a national scale infrastructure for research and experimentation in networking technologies and applications.

University of California, Santa Barbara Coordinating Node

Facilities for the UCSB Coordinating Node, personnel, and working groups will be provided by the National Center for Ecological Analysis and Synthesis (NCEAS), the Ecoinformatics Center at the Marine Science Institute, and the Davidson Library. UCSB facilities are connected via the campus inter-building backbone network that provides service over single-mode and multimode fiber at rates up to 10Gb/s.

Meetings for working groups and personnel engaged in education and outreach will be held at NCEAS in downtown Santa Barbara, CA. NCEAS has supported synthesis and collaboration among over 4000 scientists in over 400 working groups since 1995. These synthesis activities are supported by dedicated logistics staff to coordinate meetings and provide computing resources and support. Two modern meeting rooms are available, one accommodating up to 25 people and the other up to 18 people, as well as larger conference facilities. Additionally, breakout rooms are available for smaller groups. Rooms are equipped with wireless and ethernet Internet access, LCD projectors, SMART boards, and white boards. NCEAS' computing facilities include an internal high-speed LAN, dedicated computing resources for working group participants on our high-end SMP servers and compute cluster, and dedicated support from scientific computing support that specialize in organizing and managing complex data sets, designing quantitative methods for statistical analysis and modeling and selecting appropriate software to conduct these analyses.

Office space will be provided in the Ecoinformatics Center in the Marine Sciences Building. This state-of-the-art science facility was completed in the Fall of 2004. It includes high-speed networks within the building and to the campus intranet, as well as high-bandwidth connectivity to wide area networks such as the Internet. MSI currently houses the computing facilities for existing data preservation networks such as the Knowledge Network for Biocomplexity (KNB) and the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) in its dedicated server facility. DataNetONE personnel will be housed with other engineering professionals working on related projects in the Ecoinformatics Center.



Figure 4: NCEAS scientists interact closely in small to large groups, both formally and informally.

Equipment for the Coordinating Node will be housed with other digital preservation facilities at the UCSB Davidson Library. This 400 sq.ft². computing facility is cooled with an internal Liebert Deluxe System 3 capable of handling 22 tons. The room is fiber connected into the campus gigabit backbone and nine Catalyst 3570 switches (all interconnected on the back-plane) provide LAN connectivity for internal networks. Wide-area networking is provided via UCSB's CalREN2 connection, an optical link supporting multiple lambdas and data rates. CalREN2 directly exchanges traffic with various research, educational, and commercial networks, including Internet2, the Corporation for Education Network Initiatives in California (CENIC), and the Department of Energy ESN².

Oak Ridge Campus Coordinating Node

The University of Tennessee (UT) partnered with Oak Ridge National Laboratory (ORNL) to establish the Joint Institute for Computational Sciences (JICS) to advance scientific discovery and state-of-the-art computer science. JICS is housed in a new 52,000 sq. ft. building, constructed by the state of Tennessee on the ORNL campus. The JICS team includes UT-ORNL Joint Faculty appointees, JICS Research Affiliates, postdoctoral fellows, graduate students, and administrative staff. JICS Joint Faculty members hold a dual position as faculty members within a University of Tennessee department and as staff scientists within an ORNL research group. Bruce Wilson, ORNL Research Staff Member and a co-I on DataNetONE, will hold a UT-ORNL Joint Faculty appointment part-time for DataNetONE. Computing resources for DataNetONE will be purchased through the University of Tennessee and housed on the Oak Ridge campus. The Oak Ridge Campus –the joint collaboration of UT and ORNL-- provides an innovative and productive organizational structure to support advanced research and cyberinfrastructure for the nation.

The ORNL has embarked on an innovative facilities strategy of building privately owned facilities on Department of Energy (DOE) land and providing those facilities on a lease basis to DOE and other parties with a facilities need and a rationale for location on the Oak Ridge Campus. This process has triggered successful collaborations between ORNL and DOE and other entities such as NSF, NASA, UT, other



Figure 5: Computational Sciences Building.

universities, and others. One of those buildings is the Oak Ridge Computational Sciences Building (Figure 5). It has 40,000 square feet of raised floor computing space with over 10MW and growing of available power for all systems as well as requisite cooling and plenum space. The computing facility also has 7x24 operations staffing for systems and utilities; security including fire protection (on-site fire department with 2-3 minute response time), 7x24 cyber security monitoring and response, automated card-key entry and camera monitoring systems for machine room access; and an excellent local power reliability from its TVA supplier partner with multiple redundant power feeds, substations and power plant giving more than 99.999% facility power reliability.

Due to its Manhattan project heritage, the ORNL campus is large enough to have geographical diversity. In addition to the CSB located in the 5000 area, DataNetONE will also have access to additional remote computing and networking resources including some facilities located 0.5 mile away in the 1500 area creating the opportunity for disaster recovery diversity of operations within the ORNL campus. Computer hardware systems for DOE's Atmospheric Radiation Measurement (ARM) archive and DOE's

Carbon Dioxide Information and Analysis Center (CDIAC), are located in computer rooms in the 1500 area.

This Computation Sciences Building houses several high performance computing and data archive projects of relevance to DataNetONE including 2 NSF TeraGrid resource provider sites; the Neutron Science TeraGrid Gateway (ORNL- STG) (CO-I Cobb is also PI for this resource) and the new National Institute for Computational Science (NICS) track 2 supercomputer. It is also the location of the DOE's leadership computing facility (LCF) at the National Center for Computational Sciences. The ORNL-NSTG will be a DataNetONE participant assisting in data movement. The LCF and NICS each represent large, long-term, agency investments in highest end computational resources. Each will provide on the order of a peak Petaflop of Cray computing power and several hundred Terabytes of rotating storage for NSF and DOE commencing this year and next. Each facility represents a candidate large back-end storage potential partner for DataNetONE according to the strategy outlined in the project proposal.

Beyond that, the Oak Ridge computational resources also have available large tape archives that support multi-Petabyte archives currently and have expansion capacity to reach Exabytes and beyond. A High Performance Storage System (HPSS) instance is available and installed which allows direct API interface to tape storage programmatically or as part of normal center operations (Figure 6).

In addition the CSB is also the host location for many long-standing and currently active data archive initiatives including the NASA-DAAC, USGS-National Biological Information Infrastructure (NBII) and USA-NPN. Co-PI Cook and Co-I Wilson are involved in the management and operations of



Figure 6: HPSS silo.

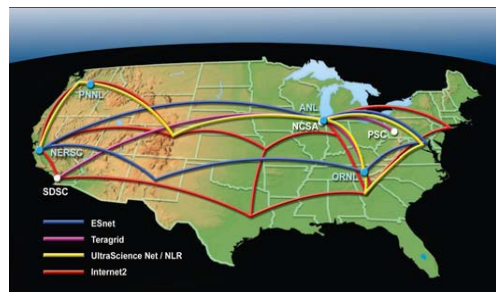


Figure 7: Oak Ridge network connectivity.

these resources. Together they represent activities that are long-standing (DAAC, CDIAC), highly utilized and important for policy decisions (CDIAC) large scale (DAAC, ARM) as well as long-term and innovative infrastructure (DACC, NBII, USA-NPN)

The Oak Ridge Campus facilities are connected with most major national and international network connections, usually at the highest bandwidth available including 10Gbs connections to NSF's TeraGrid network, ESnet, National Lambda Rail and Internet2, as well as experimental networks such as ultrascience net and cheetah (Figure 7).



Figure 8: Everest visualization wall.

Finally, the CSB is co-located with the DOE and NICS office space and meeting room that include several modern conference rooms with video-conferencing and access grid capability. In addition, there is a local state of the art visualization theater called EVEREST with a 35 Megapixel display wall and a dedicated 64-node visualization cluster (Figure 8).

The Computer Science Building physical facilities alone represent an investment well in excess of \$50 million. The facility as well as the located projects, staff, expertise, and general environment as a significant project leverage that is available to DataNetONE without additional project cost.

The University of Tennessee College of Communication's School of Information Sciences includes 3859 square-feet of office space and 801 square-feet of meeting and computer lab space. The furnished offices for faculty have late-model desktop or laptop computers many with dual monitors; 100 Mbps Ethernet and wireless access is provided by the University of Tennessee network services. Each office has a printer, and networked B&W/color printing, photocopying, scanning and fax services are available within the complex. The meeting room has a late model multimedia projector, a Polycom system, wireless access and two Ethernet ports. The student computer lab has 6 computers with Ethernet connections. The school hosts four servers including windows, Linux and Xserver for school and college use and a windows server devoted to database research.

The College's Center for Information and Communication Studies (CICS) has 538 square feet at the campus site and 1433 square feet of additional offices at the University's Conference Center Building. CICS' administrative staff includes a director, two assistant directors, and administrative personnel. CICS staff has late-model desktop or laptop computers and 100 Mbps Ethernet and wireless access is provided by the University of Tennessee network services. The adjacent college meeting room is a "Smart classroom" with a dedicated Apple computer, late model multimedia projector, wireless access and Ethernet ports.

University of Tennessee College of Communication and Information laboratory facilities include computer labs, a user experience lab, video production labs, and a video conference lab. The 612 square-foot user experience laboratory is composed of two user work station rooms, a researcher's observation room, and a reception room. Two subjects can interact with information systems simultaneously, with software control and observational equipment run by the researcher from the central observation room. The reception area provides a comfortable place for pre- and post- testing. State-of-the-art hardware and software includes cameras, computers, printers, software and a researcher/controller station enabling the collection, analysis and presentation of observational data. The software will record and interleave the keystrokes, interactions with the systems, postures, gestures, and movements of subjects. A project can be analyzed at once or the results can be exported to other data analysis software packages.

The National Evolutionary Synthesis Center (NESCent) - Duke University, University of North Carolina, and North Carolina State University

The NESCent office complex is 8,560 square-feet, including 3,362 square-feet office space and 4670 square feet of conference space and breakout areas. The furnished offices are served by the Duke telecommunications network with 100 Mbps Ethernet and wireless access to the Duke network. Networked B&W/color printing, photocopying, scanning, and fax services are on-site. All staff, as well as resident scientists, are equipped with late-model desktop computers or laptops with wide-screen monitor ports.

NESCent employs a full-time logistics coordinator for meetings and events and has standing arrangements with travel, housing and catering providers. There are three modern conference rooms each equipped with video projector and teleconferencing facility. The Center also has a portable Polycom (audio/video conferencing) system and a Symposium device that allows on-the-fly manual markup of any content visible on a computer screen. A large and configurable meeting space seating 47 people is also available, with the same facilities as the conference rooms. The conference rooms are surrounded by break-out areas with white boards. The Duke University wireless network is available to guests in all meeting spaces as well as break-out areas.

NESCent operates six load and application-optimized Apple Xserve and Linux servers for development and production of databases, web-applications, and infrastructure software. The hosting environments for production are physically separate but hardware-redundant from those for development or testing, allowing to swiftly and extensively test upgrades, bug fixes, or new features in their destination environment. NESCent also manages mailing lists, a farm of collaborative wiki sites, and a Subversion-based software source code repository. A shared file server provides more than 2TB of RAID storage for large files and databases and has enough unused disk slots available to double the capacity. The data center is supplied with supplemental air conditioning, Gigabit network connections, UPS for backup power, and a combined file/tape backup system. NESCent maintains priority access to sixteen nodes and idle-time access to several hundred other nodes on a shared high-performance computing cluster managed by Duke University. In addition to twelve high-end workstations and fourteen laptops for employees and resident scholars, the Center also maintains a shared computer lab with five workstations and six laptops with both Windows and MacOSX operating systems for shared use by Center fellows and visitors.

The Center employs a dedicated system administrator and a desktop support specialist. Together, they support a variety of operating systems (Windows, Linux, MacOSX), databases (PostgreSQL, MySQL), web and application servers (Apache, PHP, JBoss, Tomcat), and collaborative as well as productivity applications (Mailing lists, MediaWiki, WebDAV-based file shares, dotProject, MS Office, etc). An issue tracking system efficiently manages trouble ticket tracking, resolution, and prioritization.

Appendix A1. Sustainability Plans

Over the first five years of DataNetONE, annual operation costs will be tracked and future operating costs will be projected. Achieving economic and technological sustainability for DataNetONE entails an iterative and learning process that involves experimentation, adaptation, and evolution. The DataNetONE strategy is based on: seeking input and engagement from the broad stakeholder community; developing and adopting agile, robust technologies and interoperability solutions; considering the full life cycle costs for the technologies adopted; creating a strong, but flexible management structure that can evolve and scale as the organization grows; projecting future revenue streams and experimenting with a variety of sustainability approaches; and performing rigorous evaluation and assessment of the sustainable approaches that are implemented.

Long-term preservation and access to data are facilitated by the DataNetONE enterprise, which encompasses a distributed network and leverages the considerable expertise of well-established, participating organizations (i.e., archives, libraries, academic institutions, businesses, environmental observing systems and research networks, science synthesis centers, government agencies, and professional societies). The DataNetONE organizational structure encourages input throughout all levels of the organization. This begins with an External Advisory Committee (EAC) whose members provide a wealth of knowledge useful not only in reviewing and shaping the DataNetONE mission and objectives, but also in identifying best practices in the community. The DataNetONE Executive Director (ED) allocates significant time to developing a sustainable business model for the network. The collective experience of DataNetONE investigators and EAC in fund raising, meeting community needs, and awareness of past successful and unsuccessful sustainability efforts will be key to DataNetONE success.

DataNetONE supports a Working Group that is focused on Long-term Sustainability and Governance (Section 5.2.10). This group will encompass leaders throughout science and industry. Sustainability models that have emerged from the S&T community, federal agencies, and business will be sought as examples of approaches that DataNetONE could adopt. DataNetONE will evaluate the feasibility of developing a 501(c)(3) to facilitate long-term sustainability. Two workshops will be held annually to support participants in developing and implementing the DataNetONE economic sustainability plan.

The DataNetONE user community, via the DataNetONE International User Group (DIUG), will be key in supporting the sustainability of DataNetONE products and services through participation in routine community surveys and assessments; shaping DataNetONE tools, methods, and services; and active working group involvement. Both economic and technological sustainability strategies are central to DataNetONE becoming a long-term organization. These two components are different, but complementary, and require considerable investigation by DataNetONE as discussed below.

A1.1 Economic Sustainability Plan

In Japan, a company proved that an enterprise focusing on socially important issues can stay relevant for more than a millennium. The Kongo Gumi Company succeeded for 1,400 years because it picked a stable industry – temple construction – and created flexible succession policies [63]. DataNetONE aspires to similar longevity – to be an economically and technologically viable gateway to Earth observation data sets for centuries.

The DataNetONE architecture facilitates sustainability since data storage and curation are distributed to Member Nodes that fund their own participation at a level appropriate to their needs. Such an approach avoids a bloated, centralized data store that is struggling to keep up as the network grows. Furthermore, the way in which replication is managed by the Coordinating Nodes allows Member Nodes to assist each other by providing backup services for one another.

Another major component of economic sustainability is to assess long-lived enterprises, as well as existing economic and technological sustainability plans for other local, regional, national, and global research networks and CI enterprises. Various scientific and technical organizations such as the Global Biodiversity Information Facility (GBIF), Inter-American Biodiversity Information Network (IABIN), World Wide Web Consortium (WC3), and others have devoted considerable resources to developing sustainable business models. For instance, WC3 has developed a “Supporters Program” [64] and IABIN is developing a sustainability strategy [65] to support its long-term funding needs. The Long-term Sustainability and Governance Working Group will carefully review these and other efforts for their applicability to the DataNetONE organization.

In addition to constantly searching for and assessing sustainability plans from other organizations, DataNetONE's sustainability strategy features evaluation, testing, and adoption of multiple approaches from a varied portfolio of income streams including:

(1) Tiered dues-paying membership in the organization. This model is based on an association model, with different members donating according to their ability to pay (financial and/or in-kind). Some members with valuable data sets may not be able to contribute financially. Other organizations may see value in the ability to access all their data in one place, and may contribute financially. This approach requires careful management of the membership to ensure that all members are deriving value from participation, and to ensure that sufficient funds exist. We will design a model that will enable individual scientists to participate in DataNetONE at no cost, whereas organizations may pay a nominal fee based on the services that are desired (Figure 1). This model will be further explored for both short-term and long-term feasibility through the DataNetONE Long-term Sustainability and Governance Working Group, in conjunction with the DIUG, NSF, and other DataNet partners.

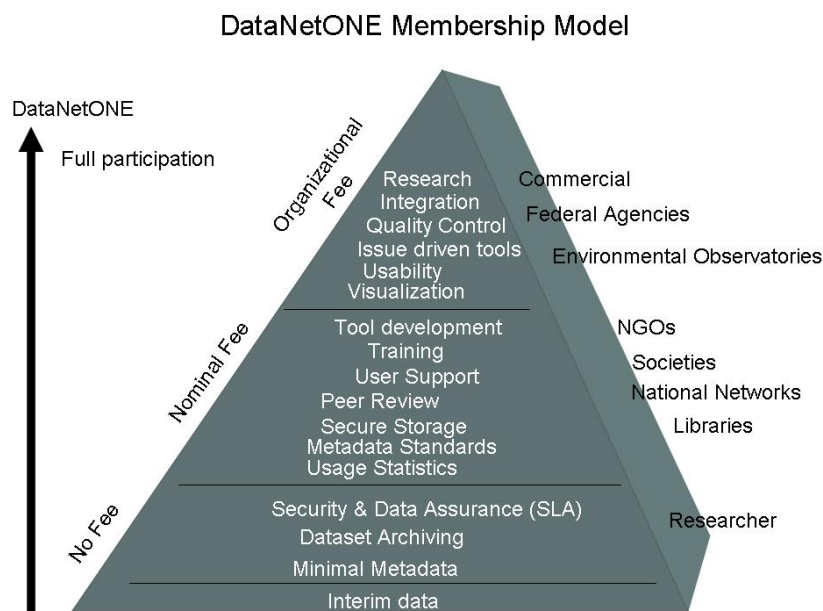


Figure 1: The DataNetONE fee-based membership model provides differing types of participation based on different user needs. Individual researchers can both access data and contribute data for archiving at no cost to themselves. Libraries, societies and similar organizations can participate as Member Nodes and utilize various services at a nominal fee. Large organizations that will need massive data storage and heavily customized services will invest through an organizational fee that allows them to take advantage of the economies of scale associated with DataNetONE.

(2) A project-driven model. This concept will derive funding from projects that are of particular value to various organizations and agencies. DataNetONE will be most successful if it can develop a multitude of applications and services that exploit the data resources that are curated within the DataNetONE network. Data curation will enhance the value of the resources by enabling persistent archiving, data synthesis, visualization, exploration, and hypothesis generation and testing. Through proper storage and management of network resources, DataNetONE will provide the foundation for the development of a suite of visualization and analysis services in support of federal, state, and private organizations' research and management needs (i.e., environmental impact statements, State Wildlife Action Plans). Additionally, DataNetONE will be at the forefront of new research in ecology where the synthesis of a multitude of variables that impact ecological processes (i.e., environmental, landscape, human demographic, weather, climate) can be linked to observations collected in DataNetONE. For example, Figure 2 illustrates observations from a variety of sources and sensors that can be linked via particular variables (i.e., location or date).

(3) An endowment built via various mechanisms. Agencies and organizations, or potentially philanthropic foundations with a long-term view of the importance of long-term data preservation will be encouraged to consider capacity grants to supplement an endowment that can serve DataNetONE for decades.

DataNetONE Services

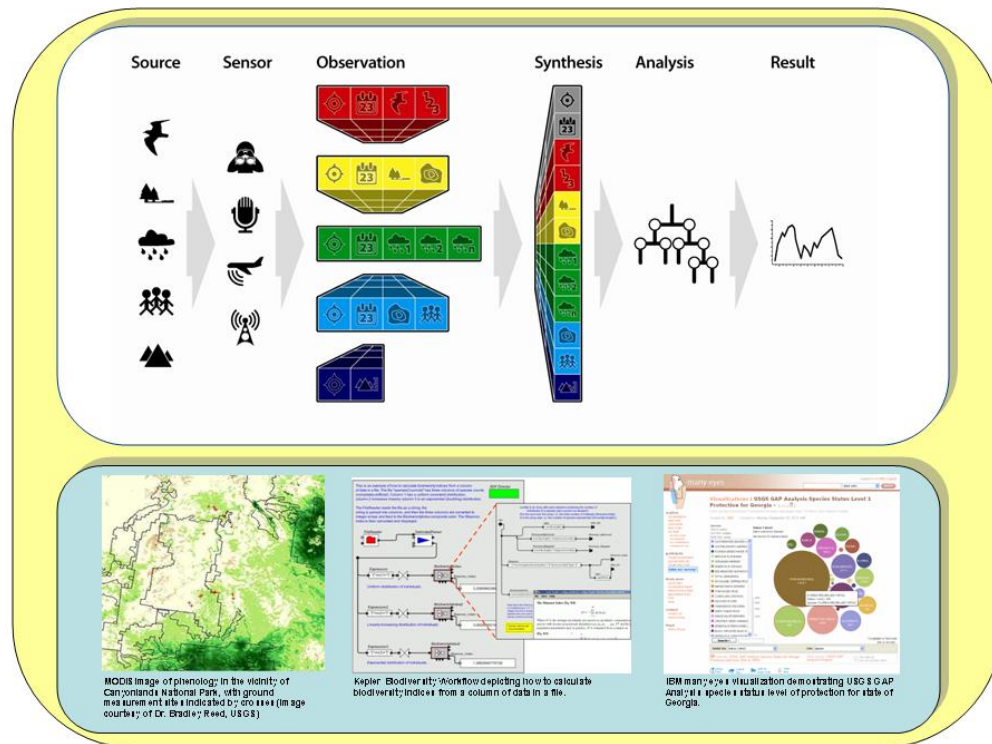


Figure 2: Examples of DataNetONE value-added services that may generate a revenue stream (e.g., geospatial data processing, scientific workflow solutions, and high-level analyses and visualizations). A variety of sensors convert environmental sources (i.e., organisms, landscapes, weather, human population density, or elevation) into observational data, which have similar and dissimilar characteristics. Based on the similar characteristics (i.e., location and date) these disparate data sources can be linked and integrated. This allows exploratory analysis to look for patterns of biological interest that lead to further analysis and the publication of results.

(4) A fee-for-service model where applications are developed or customized for a data user's specific need. In this case, users seeking specific data sets, tools, or applications may provide funds.

(5) Conference fees. One of the chief methods for increased awareness and dissemination of knowledge about DataNetONE is through annual conferences. Following this approach, a portion of the conference registration fee could be allocated to cover DataNetONE operating costs. Such an approach has been adopted by many professional societies to generate revenues.

(6) Administrative fees from grants. The potential exists for DataNetONE products and services to be utilized to support new research efforts. DataNetONE will investigate the applicability of such projects contributing modest funds to support DataNetONE activities related to infrastructure improvements and sophisticated tool development.

(7) A Wikipedia model with contributors offering time and resources. DataNetONE will also seek input from venture capital and organizational sustainability experts participating in the DataNetONE Long-term Sustainability and Governance Working Group (cf. Sections 4.2 and 5.2.10). Ongoing sustainability may partially depend on the willingness of volunteers to contribute code or other services.

(8) Matching funds, in-kind services. Matching funds not only allow DataNetONE funding to go further, but also help support community buy-in, involvement, and long-term sustainability of DataNetONE activities. Various models for leveraging the significant resources of existing DataNetONE partners, including facilities, staff, hardware/software, etc., will be investigated in the first year of the project. It is anticipated that significant contributions by DataNetONE partners, especially in the areas of facilities and long-term commitment of staff time, will be devoted to DataNetONE by the organizations involved.

A1.2 Technological Sustainability Approach

Achieving technological sustainability presents unique opportunities and challenges for a DataNet organization. DataNetONE will be based on distributed systems architecture, with redundant service providers, metadata repositories, and data storage nodes. The Core Cyberinfrastructure Team (CCIT) will monitor, evaluate, and test evolving systems, technologies, and standards in core and member nodes.

Technological sustainability approaches to be evaluated and implemented in DataNetONE include:

(1) Distributed systems and replication of services architecture. The use of a distributed systems architecture allows DataNetONE to capitalize not only on the technological expertise of its partners, but also leverage existing infrastructures of DataNetONE Member Nodes. Replication of various services and functions throughout the network ensures that multiple copies of data and services exist and that heterogeneous environments can be supported by the DataNetONE architecture. This results in wider adoption by the community, reduction in hardware/software costs, and demonstrates the tremendous value of DataNetONE participation. Ultimately, participation and system reliability increases because of low barriers to participation by users.

(2) Working Group and user-centered design approach. DataNetONE considers users an integral component of the economic and technological sustainability strategy. DataNetONE activities are primarily carried out through Working Groups. This presents tremendous opportunities for sustainability due to the extensive stakeholder involvement as well as leveraging of participant expertise and resources.

(3) Adoption and participation in standards. The adoption of existing content and technological standards ensures DataNetONE can replicate DataNetONE-managed data, systems, and processes. Significant leadership by DataNetONE partners exists that will be leveraged in support of DataNetONE standards activities (Appendices A4 and A5).

(4) Open source non-proprietary development and archival solutions. The adoption of “open-source” methodologies is paramount to DataNetONE sustainability as it relates to the incorporation of existing efforts (i.e., Lots of Copies Keeps Stuff Safe [LOCKSS], persistent identifiers and resolver schemes (LSIDs, ARKs), format validation (JHOVE; see [66]), obtaining community buy-in, providing a wide range of tool developers/maintainers, and ensuring DataNetONE products and services are available to a broad range of users for years to come.

(5) Identifying base level functions of DataNetONE. DataNetONE will work to identify minimal and optimal levels of functionality that would be required to sustain DataNetONE activities after year 11. In its simplest form, DataNetONE will retain a binary copy of the data set, identify a minimal set of tools required to support its preservation, and develop key conversion/translation tools to support its discovery and availability. Optimally, rich metadata and comprehensive documentation would be present to support multiple long-term data uses.

(6) Educate stakeholders about best practices. DataNetONE will be a catalyst in the community to foster change in earth science through its services and various incentives offered for participation. DataNetONE base and value added services will aid in this cultural transformation of data management, including long-term curation. DataNetONE working groups along with the Associate Director for Community Engagement and Outreach will lead this endeavor.

(7) Development of QA/QC protocols and standards. DataNetONE understands the importance of maintaining the original format, content, and structure of data entrusted to DataNetONE. Data integrity checks, data migration services, check-sum techniques, and other necessary technologies will be adopted and developed when necessary.

The DataNetONE sustainability approach will clearly be an iterative and evolving process. DataNetONE believes that through its strong management structure, long-term commitment by its partners, and broad community engagement a long-term sustainability model can and will be developed.

Appendix A2. Management Plan

DataNetONE is a flexible, learning organization that can grow and evolve over time to meet changing scientific and cyberinfrastructure needs, as well as adapt to different sources and levels of funding. The management plan includes three components that are discussed below.

1. Organizational Structure and Key Leadership Positions

DataNetONE has a logical management and reporting structure that promotes strong oversight and accountability, optimizes internal and external communications, and can easily transition into an independent corporation (e.g., 501 (c)(3) and Limited Liability Corporation) (Figure 1). The DataNetONE management structure is also designed to support broad community engagement, be responsive to ever-changing technologies, and foster incubation of research concepts.

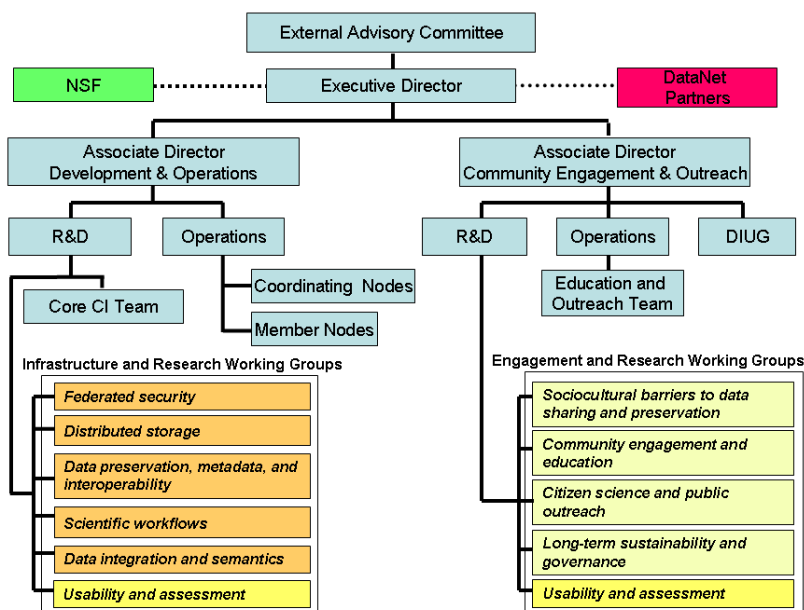


Figure 1: The DataNetONE organization chart. Bold lines depict management hierarchy within DataNetONE and dashed lines show ties with NSF and other DataNet projects. The DataNetONE International User Group (DIUG) is the worldwide community of Earth observation data authors, users, and diverse stakeholders.

Key leadership positions created at the inception of DataNetONE are defined below. As the organization matures, staff can be added or removed. The DataNetONE Office will be located in Albuquerque, NM. Because DataNetONE is principally a distributed organization, we anticipate that some office functions may ultimately be subsumed by DataNetONE partnering organizations or out-sourced to commercial enterprises (e.g., Administaff, Inc.) if the virtual organization assumes independent 501c3 status.

External Advisory Committee (EAC). The organization will be overseen by an independent EAC that ensures that DataNetONE is fulfilling its mission, serving its user community, and establishing and implementing a sustainable business plan. The nine members of the EAC will have a range of internationally-recognized expertise relevant to DataNetONE—scientific and technical, scientific and cyberinfrastructure enterprise management, library and archival practices, economic sustainability, data provider / data user, policy and decision making, and leadership in relevant professional societies, government, and industry. The Co-PIs will initially comprise the EAC and they will establish a charter describing EAC member selection, mission, operations, and reporting. Term limits will be established to ensure stability and to establish a rotation of one-third of the members each year; thereafter, EAC members will be appointed for 3-year renewable terms.

Additions to the initial EAC will be selected through a nomination and selection process managed by the Co-PIs. Nominations will be solicited from the broad scientific community. The Co-PIs and Co-Is will prioritize candidates in four areas—science, cyberinfrastructure, management/sustainability, and the

library community. A final slate of EAC members will be coordinated with input from NSF to ensure appropriate diversity (gender, racial, ethnic, geographic, and institutional) and adequate representation of the four foci. Once established, the full 9-member EAC will nominate replacements in consultation with the ED, ADs, and NSF.

The EAC will convene twice annually for meetings of 1-2 days. A Chair elected by the EAC for a 2-year term will lead meetings and prepare reports to be made publicly available. The EAC participates in recruiting and selecting the Executive Director (ED) following University of New Mexico (UNM) hiring procedures, annual performance reviews of the ED and regular reviews of DataNetONE finances. The EAC can easily transition into a Board of Directors should DataNetONE decide to incorporate as a 501c3 organization. In this case, bylaws will be developed and the Board will develop a succession policy maintaining the independent status of the Board and ensuring broad community representation.

Executive Director. DataNetONE will be administered and managed by an ED who is responsible for all technical, management and budget issues and is an employee of UNM. The ED reports to the EAC and is the principal point of contact with NSF and with other DataNet partners. The ED will also develop, evaluate, and implement strategies for economic sustainability of DataNetONE.

The ED will have a Ph.D. in science or information technology; demonstrated expertise with the earth observation and cyberinfrastructure communities; ten or more years experience in managing large, complex, distributed scientific enterprises; and strong leadership and communication skills. Ideally, the ED will have significant experience in fund-raising (e.g., grants-writing) and organizational development.

Recruitment and Selection of the Executive Director. A subcommittee of the EAC will initiate an international search with advertisements in *Science* and *Nature*. The Search Subcommittee will solicit applications, contact references, and recommend three candidates for consideration by the full EAC. The EAC will conduct interviews and recommend the top candidate to NSF and UNM, which will employ the ED. Until a permanent ED is hired, William Michener (DataNetONE PI) will serve in this capacity.

Associate Directors. The Associate Director for Development and Operations (AD D&O) will oversee development and implementation of DataNetONE architecture, computer science research agenda, and technological evolution through the activities of a relevant subset of the working groups and the Core Cyberinfrastructure Team (CCIT). The AD D&O is responsible for establishing the CCIT work plan, providing management and oversight of the developers, and implementing new technologies at DataNetONE nodes. An Associate Director for Community Engagement and Outreach (AD CE&O) will implement outreach and education activities through DataNetONE member organizations and coordinate relevant working groups. The AD CE&O will engage the community through the DataNetONE International User Group (DIUG), which is made up of earth observation data authors, users, and holders (e.g., students, educators, researchers, libraries, data centers, professional societies, agency representatives, policy-makers, and the general public). DIUG will meet annually to identify the evolving technical challenges and opportunities for advancing education, research, and policy through the use of DataNetONE data products, tools, and services. In addition, the AD CE&O leads community development, education, and outreach activities associated with the DataNetONE nodes and the relevant subset of working groups. Both ADs will participate in relevant working group activities to ensure adequate communication and integration of activities and results.

Recruitment of the ADs will commence as soon as funds are awarded. The interim ED (Michener) will initiate the process following UNM hiring procedures. DataNetONE Co-PIs will serve on the AD selection team. Bruce Wilson (ORC) and Matthew Jones (UCSB) will serve as co-leaders of the CCIT and will share responsibilities of the AD D&O until the position is filled. Co-PI Stephanie Hampton and Co-I Vivian Hutchison will similarly share responsibilities of the AD CE&O until that position is filled.

2. Reporting Relationships, Oversight, and Accountability Mechanisms

The Chair of the EAC provides input on the annual performance evaluation of the ED to the UNM Vice President for Research. The ED serves as the principal point of contact with NSF, the EAC, and the leadership of other DataNet partners. The ED communicates with appropriate NSF Project Directors, and submits annual reports to NSF and semiannual reports to the EAC. The ED, the two ADs and one of the CCIT Co-Leaders represent DataNetONE at the NSF-DataNet Partner meetings.

The ED supervises the two ADs who report weekly on progress, challenges and project risks. The AD CE&O oversees office staff and coordinates and communicates with the Education and Outreach Team, the DIUG, and four of the Community Engagement Working Groups. The AD D&O oversees activities of the CCIT and coordinates and communicates with node personnel that comprise DataNetONE operations

and five of the Infrastructure and Research Working Groups. Both ADs will interact closely with the Usability and Assessment Working Group to ensure products and services fully meet user needs.

The CCIT and each of the Working Groups will have two co-chairs to ensure that the leadership work load is shared and to guarantee continuity in communication and coordination. Each working group and the CCIT annually propose a work package and timeline for completion of work products to the respective AD. The co-chairs and ADs will prepare quarterly reports documenting accomplishments, plans, and challenges to maximize integration of efforts throughout DataNetONE.

3. Communication among DataNetONE Members and with the Community

DataNetONE supports numerous face-to-face meetings that are central to project planning, work activities, and oversight as well as community engagement, outreach, and education (Figure 2).

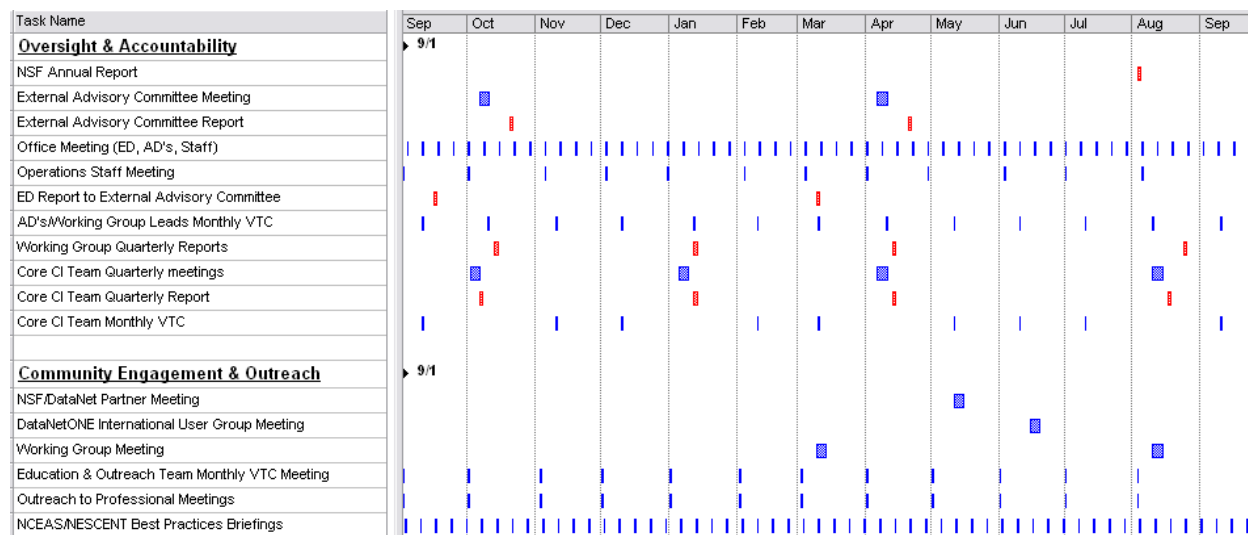


Figure 2: Annual schedule of meetings and reporting. Planned meetings include those that support interactions among: (1) DataNet partners and NSF; (2) the DataNetONE EAC; (3) Working Group participants; (4) members of the CCIT; (5) registered participants in the DataNetONE International Users Group; and (6) outreach visits whereby DataNetONE-associated educators and outreach specialists offer workshops and training sessions for the broad community (e.g., professional society meetings).

In addition to the DataNet partners meeting required by NSF, DataNetONE is designed so that other DataNet partners can participate in Working Groups as well as the annual meeting of the DIUG. Co-PI Mike Frame will routinely apprise an inter-agency cyberinfrastructure Working Group (BioEco—comprised of representatives from NASA, USGS, NSF, EPA, USACOE, USDA) of DataNetONE activities. Specific communication with State Agencies will occur through the Organization of Fish & Wildlife Information Managers (OFWIM) via DataNetONE collaborator Vivian Hutchison, current OFWIM President.

The DIUG is a membership-based organization with a Chair and Vice-Chair that meets annually to provide community feedback. The DIUG will propose and prioritize action items (future tool and service development) that the ADs will task to the CCIT and Working Groups. Working Groups will develop research products (e.g., publications, tools, and standards) or a detailed set of recommendations that address specific infrastructure, research, and engagement activities originating with the DIUG.

The Usability and Assessment Working Group is the interface between operations and community engagement and is dedicated to developing surveys and other instruments that provide feedback to DataNetONE. The CCIT is responsible for setting specifications for services, supporting common network services and programming needs, and implementing tools and services developed by the Working Groups. Communication is vital in this distributed and feature-driven product development model. In addition to quarterly face-to-face meetings, CCIT members will videoconference monthly and will rely heavily on informal communication mechanisms such as instant messaging for daily communication. The web-based DataNetONE Collaboration Space will support conferencing so Working Group and team members can meet virtually in between face-to-face meetings. The DataNetONE Portal will serve as the primary gateway to data holdings and information (e.g., news and reports) about the organization.

Appendix A3. Cyberinfrastructure Capabilities

Cyberinfrastructure capabilities include four categories that are discussed below.

1. System Architecture and Services. DataNetONE will be a robust, distributed network comprised of a small set of replicated *Coordinating Nodes* and a much larger set of *Member Nodes*. Coordinating Nodes choreograph services that help users store, discover, and retrieve data while Member Nodes archive data using various standards and formats. The Coordinating Nodes guarantee replication and preservation of data objects, authorization and authentication, and data discovery. They ensure that any archived data object is copied to at least two other geographically and organizationally distinct Member Nodes and are responsible for authentication and authorization for any operations by users or services on DataNetONE. Member Nodes minimally act as simple data storage devices supporting CRUD (create, read, update, and delete) operations on digital objects, with each object assigned a persistent unique identifier using one of several well known identifier schemes (e.g., LSID [17], ARK [18]). Member Nodes would optionally contain a local metadata index and support services such as resource discovery and optional services such as analysis and visualization as needed.

Coordinating Nodes. The network is designed to support persistence of scientific products throughout the data lifecycle via geographic redundancy, failover, and high-availability. The requirements of Coordinating Nodes and Member Nodes differ significantly, with the Coordinating Nodes designed and located to maximize uptime, connectivity, and security. These nodes play a critical backbone role and will be virtualized in order to support failover and load-balancing across the redundant geographic locations. Client systems that communicate with the Coordinating Nodes will be dynamically routed to one of the replicas for the duration of a transaction, and these service transactions may span multiple calls to the DataNetONE Service Interface to promote consistency and ease of replication. Replication technologies in use in relational data management systems and in data management middleware (e.g., Metacat, Storage Resource Broker) will be robust even in the event of extended system and network outages.

Large distributed systems in the environmental sciences like DiGIR and TAPIR have demonstrated the effectiveness of local data management when local expertise is needed to curate data holdings. However, these systems have limitations in terms of search performance and query consistency when large numbers of nodes are needed to perform a query. Consequently, DataNetONE Coordinating Nodes will support critical network-wide search services by maintaining a comprehensive metadata index for data from all Member Nodes. This architecture will provide consistent results and rapid responses due to efficient, centralized metadata caching from the redundant, high-availability set of Coordinating Nodes. Member Nodes have less stringent reliability requirements, since any archived object is replicated in several other Member Nodes; the chances will still be very high that a desired object can be retrieved even if one or several Member Nodes are offline. Member Nodes that are offline for long periods are detected by the Coordinating Nodes heartbeat and health services, which will trigger contingency plans for assuring adequate data replication in place of the affected Member Node.

DataNetONE Service Interface. The services provided by the Coordinating Nodes and by the Member Nodes, as well as all interactions between nodes and client tools in the Investigator's Toolkit, will be developed as part of a shared Service Interface that covers Network Services, Federated Identity and Authorization Services, Object Management Services, Preservation Services, and Discovery and Usage Services (see Section 3). This will allow sites with existing infrastructure to adapt their systems to conform to the service interface, and thereby forego the cost of deploying altogether new systems, although this will be an alternative as well. The service interface will extend existing services like the EarthGrid [23] web service interface that already operates across several common data systems in the environmental sciences, and various services developed in the Grid community (e.g., the Globus Notification service).

Additional services may be layered upon the DataNetONE infrastructure and can offer value added capabilities such as data analysis and visualization operations. The modular design of DataNetONE allows these services to be operated on Coordinating or Member Nodes. To contribute as a Member Node, participants may run their own hardware of different capacities, or may choose to outsource to a commercial provider. Such flexibility is expected to minimize the cost of participation in DataNetONE.

Member Nodes. A critical number of Member Nodes is required to provide scalability. Within DataNetONE, data object storage is delegated primarily to the Member Nodes. Member Nodes implement a reduced number of services compared with the Coordinating Nodes, the main difference being that Member Nodes are not expected to maintain a complete copy of all DataNetONE metadata or implement network-wide services. Member Nodes are similar to the nodes of the LOCKSS network [47].

Member Node creation will be possible through a number of different scenarios from manual installation from DataNetONE source code through to the provision of a Linux distribution in both virtual machine and installation CD format. The Member Node software stack will be hardware independent and designed to scale according to the available physical resources. Once operational, a DataNetONE Member Node will register with a Coordinating Node, which will determine basic functional parameters of the Member Node (bandwidth, storage, and processing) and add it to the replication scheduler, which will then start replicating existing content to the new node based on provisioning policies established in the Member Node Service Agreement. Space will be maintained on Member Nodes to ensure adequate performance of local users of the Member Node. This proportion will be dynamically allocated, so that a Member Node administrator can request additional storage for an upcoming experiment or make more available to the DataNetONE network. Each Member Node will operate under a Service Level Agreement with DataNetONE that will specify contingencies such as data disposition if the Member Node goes offline for an extended period or merges with other nodes.

Investigator Toolkit. Community engagement is critical to the success of DataNetONE and will be achieved by creating an Investigator Toolkit that integrates with researchers' existing analysis and modeling practices. Software design will be responsive to user requirements for data archiving and will provide added value that makes it easier for scientists to achieve their research goals. These tools must stimulate cultural shifts in the data archiving habits of researchers by providing efficiency gains for practicing researchers. The Investigator Toolkit will support both researchers directly and research libraries that support scientists. Tools will assist in data collection and ingest, and help DataNetONE to provide high quality end-to-end data services by engaging early in the data generation process. We will also provide interfaces for accessing DataNetONE services within desktop analysis software, including common scientific frameworks such as Matlab and R. The system would also provide a browse, search and retrieval functionality that would incorporate DataNetONE services into current and emerging digital library services for students and faculty. To support search and retrieval of DataNetONE content in combination with other library content, for example, the system would provide interfaces for next-generation faceted search interfaces and more traditional federated search systems. The DataNetONE team will conduct research with academic and research librarians and library users to identify the necessary software requirements.

Early versions of the toolkit will be based on existing community tools that are adapted to use the DataNetONE Service Interface. These will include various data and metadata editors (Morpho), workflow engines and repositories (Kepler, myExperiment), analytical tools (Matlab, R, Sage), and grid portals. In addition, we will provide an archival tool that will operate similarly to Jungle Drive [67] or Apple's Time Machine [68] that includes the addition of metadata and, where possible, the data in a non-proprietary format. Such automation, or "single button archiving," will dramatically reduce the hurdles for effective use of data archives. By providing a DataNetONE client software library, it is expected that DataNetONE archives will also be integrated into common scientific applications. With these tools, a researcher can easily publish a data set, Kepler workflow, or R package to the DataNetONE archive and make it, along with associated metadata and references to other datasets, available to other researchers.

2. Hardware and Facilities Infrastructure. The first three Coordinating Nodes will be established at ORC, UNM, and UCSB. The ORC Node will use facilities at Oak Ridge National Laboratory and the ORNL-UT Joint Institutes (Joint Institute for Computational Science (JICS), and the ORNL-UT Joint Institute for Biological Science (JIBS)). The UNM Node will utilize facilities associated with the New Mexico Computing Application Center and Intel Corporation. The UCSB Node will use facilities at the National Center for Ecological Analysis and Synthesis, the California Digital Library/UCSB Davidson Library, and the Marine Science Institute.

At ORC, DataNetONE will utilize the physical and networking facilities at JICS for the Leadership Computing Facility (LCF). The LCF has over 40,000 ft² of space appropriate for high-end computing. These facilities are currently served by two independent 161 kV supplies from a recently constructed TVA substation, with a third 161 kV supply source due to come on-line in June 2008. The LCF currently has 15,000 tons of cooling capacity on-line, with the capacity for additional cooling as the need arises. The Oak Ridge Campus is connected to every major research network at rates of 10 gigabits per second or greater using optical networking equipment owned and operated by UT-Battelle (the LLC that operates ORNL for the Department of Energy), which has the capability of simultaneously carrying either 192 10-gigabit per second circuits or 96 40-gigabit per second circuits and connects the LCF to major networking

hubs in Atlanta and Chicago. Currently, only 16 of the 10-gigabit circuits are committed to various purposes, allowing for virtually unlimited expansion of the networking capability. As part of this proposal, we will expand the wide area connectivity to 10 × 10 gigabits per second. Currently, the connections into ORC include TeraGrid, Internet2, ESnet, and Cheetah at 10 gigabits per second as well as UltraScienceNet and National Lambda Rail at 20 gigabits per second. JICS operates the Cheetah research network for NSF and ORNL operates the UltraScience Net research network for DOE. When the need for storage rises beyond the 50-100 TB range planned for the seed storage in the Coordinating Nodes, the LCF can also serve as a data storage facility. The ORC node will be able to provision storage through the High Performance Storage System (HPSS) operated by ORNL. This installation currently provides over 5 PB of storage, and can be scaled to over 750 PB.

The UNM DataNetONE Coordinating Node will be co-hosted with the New Mexico Computing Applications Center (NMCAC) at the Intel Corporation fabrication facility in Rio Rancho, New Mexico. Intel Corporation has operated their New Mexico facility since 1980 and opened the “Fab 11X” production unit in October 2002 as part of their 4 million square feet campus. The NMCAC is home to the world's 3rd fastest supercomputer (2008), a 172 teraflops Altix® ICE system developed by Silicon Graphics, Inc. (SGI) and named “Encanto” (enchanted), with 14,336 Intel Xeon® processors, 28 Terabytes (TB) of memory, and 172 TB of online storage. The NMCAC system is supported as an integral part of Intel's Data Center infrastructure, which includes year-round continuous operational support (conditioned power and environmental control) and fiber-optic connectivity to the National Lambda Rail (NLR) network. The NMCAC provides 24x7 on-site personnel for immediate system maintenance, which ensures minimal “down-time” of all managed infrastructure. The shared facility with NMCAC/Intel will optimize functional requirements of the DataNetONE node by taking immediate advantage of the redundant and fault-tolerant environment required by NMCAC/Intel operations.

At UCSB, the Davidson Library will house the DataNetONE Coordinating Node equipment in its 400 sq.ft. computing facility that currently hosts other long-term digital preservation programs, such as the Alexandria Digital Library and the Map and Imagery Laboratory. The server room is cooled with an internal Liebert Deluxe System 3 capable of handling 22 tons. The room is fiber connected into the campus gigabit backbone and has nine Catalyst 3570 switches (all interconnected on the back-plane) providing LAN connectivity for internal networks. Wide-area networking is provided via UCSB's CalREN2 connection, which is an optical link supporting multiple lambdas and data rates. CalREN2 directly exchanges traffic with various research, educational, and commercial networks, including Internet2, the Corporation for Education Network Initiatives in California (CENIC), and the Department of Energy ESNet. The facility houses many servers and compute clusters that support the Map and Imagery Library, the CDL, and other library services. The facility serves as one of the core REDDNET sites on the west coast, committed participants in Logistical Networking for high-performance distributed storage over WANs. Systems are power connected through smart UPSs, are monitored 24/7 with Nagios, and may be remotely managed via KVM over IP. System backups are managed with Atempo Time Navigator software using 8 Sony SAIT tape drives in a Qualstar library.

Within the Coordinating Nodes, DataNetONE will use both server and storage virtualization based on the Xen hypervisor operating system paired with a set of tiered iSCSI storage arrays, an architecture previously tested at the ORNL DAAC. Additional storage can be provided by HPSS and by other storage providers, including purchased storage through, e.g., Amazon S3 and Google Scientific Storage.

Member Node Infrastructure. During the early development phase of the project, Member Nodes will be installed initially at the USGS Center for Biological Informatics and the University of Illinois-Chicago library, as well as various test and development sites to determine the useful range of hardware requirements (processing, bandwidth, storage) necessary to provide an effective Member Node installation. These initial Member Nodes will cover the range of expected clientele from researchers with simple data archive requirements to researchers requiring highly interactive use of the network for merging existing and new data sets or performing detailed analyses on archived data sets. Initial requirements for a Member Node may be as low as a basic Linux workstation or server with 0.5 GB storage and dual-core processor. After three years, DataNetONE is expected to have Member Nodes operational for at least the following institutions: NBII, Denver; GBIF, Copenhagen; IABIN, Panama; University of Illinois-Chicago Library; KU library – Lawrence; UT library – Knoxville; University of California Digital Library – Berkeley, CA; USGS, Lafayette, LA; Atlas of Living Australia – Canberra; NPN—Tucson, AZ; NESCENT – Raleigh, NC; Cornell Lab of Ornithology, Ithaca, NY; ESA, Washington DC; SanParks and SAEON, South Africa; eScience, Edinburgh, Scotland; and TERN, Taiwan.

3. Cyberinfrastructure Personnel and Help-Desk Services. The success of DataNetONE depends as much upon its personnel and organizational model as its hardware infrastructure. DataNetONE personnel represent a multidisciplinary team with expertise in earth science disciplines, library science, computer science, and cyber-infrastructure. The DataNetONE core staff will include leading edge researchers in computer and information science and engineering (CISE), as well as personnel with experience managing distributed science infrastructure.

As outlined in Section 4, the Associate Director for Development and Operations (AD D&O) and the Core CI Team (CCIT) will provide the overall architectural vision for DataNetONE and design and direct the implementation of DataNetONE systems with participating organizations and other DataNet Partners. The CCIT collaborates with the working groups by, for example, participating in working group research and implementing results of this research. CCIT co-leads will provide technical leadership while recruiting the AD D&O, who will supervise project engineers and system and network administrators. Software engineers will design, implement, test, deploy, and support DataNetONE systems. System administrators at the Coordinating Nodes, in conjunction with existing systems staff at the sites, will form an integrated systems team that provides 24/7 system coverage across the network for daily operations.

For user support, DataNetONE will create a comprehensive Help Desk that initially leverages the existing services in the ORNL DAAC and the LTER Network Office, which will allow coverage across the major US time zones and incremental help desk capacity growth. In addition, we will develop support services through Member Nodes via library services to support use of digital collections, including multi-institutional virtual reference services and scheduled face-to-face and on-demand online synchronous and asynchronous instructional programs. The integrated DataNetONE-wide support process will be based upon open and consistent help desk systems, including ticketing systems for issue, task, defect, and feature tracking systems (e.g., RT, Bugzilla) and knowledge bases.

4. Strategies for Agility in a Rapidly Changing Technological Landscape. After the initial installations at Coordinating Nodes, hardware will be on a three year refresh cycle, with components replaced or augmented as necessary to ensure overall system reliability and capacity (storage, bandwidth, processing) requirements are met. Past trends have shown a continuing increase in available storage capacity per unit cost and it is reasonable to expect this trend will continue for at least several more years. Thus, for a constant spending rate, capacity and processing will continue to increase, providing a baseline for continued increase in capacity for DataNetONE nodes. However, the amount of information to be archived is expected to increase significantly as new streams of data come online and researchers continue to work with larger data sets. Since the majority of the data is held by Member Nodes, increased capacity will be obtained by attracting additional participants to the DataNetONE system, allowing the entire infrastructure to grow organically to meet the demands of the community.

The use of virtualization technologies will create flexibility in hardware and software upgrades and migration. For example, to replace the hardware of a virtual machine (VM) host, the running VMs can be transferred to a different hardware host, the original upgraded or replaced, then the VMs migrated back to the new host all with little or no disruption to the operation of the data center. The use of shared iSCSI storage devices makes addition of capacity a simple process and has the benefit of minimizing the downtime associated with significant software upgrades to a VM (e.g. kernel replacement). In this case, the new VM can be created and configured to utilize the same shared iSCSI storage space as the VM to be replaced, and then brought online as the original is taken off. Users will see minimal disruption in connectivity during the process. Virtualization also enables a degree of hardware and vendor independence, which improves the negotiating power of DataNetONE centers when purchasing or upgrading hardware. Such an approach also makes installation of additional DataNetONE centers relatively straightforward, as anything from a single server with internal storage to a managed cluster with massive attached storage can participate as a DataNetONE node. Hence DataNetONE may take advantage of unused storage and processing capabilities at a project or institution by providing a simple VM configuration that provides the critical capabilities of a Member Node. The software will automatically synchronize and participate in the overall DataNetONE network yet be easily decommissioned or repurposed without significant overall consequences to the DataNetONE network. Such organizations would benefit from reduced latency when accessing frequently utilized datasets (i.e., something akin to local caching of data sets and perhaps processing capability).

Appendix A4. Key Personnel

Suzie Allard is Assistant Professor at the University of Tennessee School of Information Sciences. Her research focuses on the way scientists and engineers use and communicate information using both informal and formal channels, and how these communication processes influence the data preservation process at the point of data creation. Allard brings expertise in social processes and data preservation.

Paul Allen is Assistant Director of the Information Science Program at the Cornell University Lab of Ornithology and has been architecting software systems for collecting, archiving, and presenting biodiversity observations over the Internet, especially focused on amateur users and citizen scientists.

Peter Buneman is Professor of Database Systems in the School of Informatics at the University of Edinburgh, UK. His research focuses on databases, programming languages, data provenance, archiving, and annotation. He brings experience in data preservation and interoperability standards.

Randal Butler is Co-Director for the National Center for Supercomputing Applications Directorate of CyberSecurity. Butler is experienced in leading and/or collaborating on large, complex R&D and cyberinfrastructure projects including the NSF Network for Earthquake Engineering and Simulations NEESgrid. He brings expertise in federated security design and implementation.

John Cobb is the Oak Ridge National Laboratory local Principal Investigator for the TeraGrid and was part of the Spallation Neutron Source project as CIO, cyber-security officer, and IT staff. Cobb has experience in the application of scientific computing and supercomputing for scientific research, including project management skills necessary for the success of large collaborative projects. His expertise in the area of distributed file management is central for DataNetONE.

Robert Cook is a Distinguished Research Staff member in the Environmental Sciences Division at Oak Ridge National Laboratory. Cook is experienced in large interdisciplinary projects, including being the Chief Scientist at NASA's ORNL Distributed Active Archive Center (DAAC) for the past 10 years. Cook will work with the Community Engagement and Education working groups.

Patricia Cruse is the founding director of the California Digital Library's Digital Preservation Program. She works collaboratively with the ten University of California libraries to develop strategies for the preservation of content that is important to the research, teaching, and learning mission of the University. She will apply her skills in the area of long-term sustainability and governance.

David De Roure is a Professor of Computer Science in the School of Electronics and Computer Science at the University of Southampton, UK where he leads the Grid and Pervasive Computing research activities. His research is focused on the design of distributed computing systems, including grid, pervasive and future computing systems, and with emphasis on data and computation.

Ewa Deelman is a Research Assistant Professor at the USC Computer Science Department and a Project Leader in the USC Information Sciences Institute. She is leading the Pegasus project, which designs and implements workflow mapping techniques for large-scale workflows running in distributed environments.

Clifford Duke is the Director of Science Programs for the Ecological Society of America. He has expertise in ecological science and risk assessment, environmental project management, and policy analysis. He brings a broad understanding of data sharing needs in the life sciences based on his leadership of ESA's Data Sharing Initiative.

Michael Frame is the Director of Research & Technology for the U.S. Geological Survey National Biological Information Infrastructure. Frame is responsible for geospatial, informatics, standards, metadata, and technology operations at the Center for Biological Informatics and within the NBII Program. He brings a federal R&D perspective of operating a distributed data center like the NBII.

Carole Goble leads the Information Management Group at the University of Manchester, UK. Goble is a founding leader of the UK's e-Science activity and is the Director of the myGrid project, which produces the Taverna workflow workbench and the myExperiment social network system for e-Scientists. She co-leads the UK's Open Middleware Infrastructure Institute-UK. Goble will provide expertise in the area of scientific workflows.

Stephanie Hampton is Deputy Director of the National Center for Ecological Analysis and Synthesis. Hampton facilitates training, outreach, and interdisciplinary collaborative research by hundreds of scientists each year who intensively use highly dispersed heterogeneous data.

Donald Hobern has until recently been working as an information technologist for the Global Biodiversity Information Facility developing its technical infrastructure for integrating biodiversity information at the global level. He has been active in standards development and in promoting the use of shared open architectures. He is now Director of the Atlas of Living Australia project and Chair of the Taxonomic Databases Working Group.

Peter Honeyman is a Research Professor at the School of Information and the Scientific Director of the Center for Information Technology Integration at the University of Michigan. For the past 15 years, he has been leading research in distributed file system design and implementation. Honeyman will focus his talent on areas of distributed data storage and wide-area file systems as it relates to data storage, transfer, and replication within DataNetONE.

Jeffery Horsburgh is a Research Engineer at Utah State University. He is currently working on the CUAHSI Hydrologic Information System project, developing cyberinfrastructure for Hydrology and Environmental Engineering in support of environmental observatories. Horsburgh brings experience in developing and deploying cyberinfrastructure for data management, visualization, and analysis, for the national WATERS network of environmental observatory test beds.

Vivian Hutchison is the Metadata Coordinator for the USGS National Biological Information Infrastructure, coordinating the creation and dissemination of geospatial and biological metadata records on a national and international scale by conducting metadata workshops, participating on standards development committees, and overseeing the implementation of a metadata Clearinghouse. She brings significant experience in outreach activities and metadata management application.

Matthew Jones is the Director of Informatics Research and Development at NCEAS at UC Santa Barbara. He focuses on the management, integration, analysis, and modeling of heterogeneous data, through programs such as the Knowledge Network for Biocomplexity, a long-term data archive of over 15,000 environmental data sets, and Kepler, an open-source scientific workflow system. He brings experience in metadata standards, data management software, and scientific workflow systems.

Steve Kelling is the Director of the Cornell Lab of Ornithology's Information Science (IS) Program. He has successfully developed and managed eBird, an Internet application that collects 300-500 thousand observations monthly from a network of citizen scientist participants. Kelling has also led the creation of Birds of North America Online, a comprehensive author, editor, and community Internet environment built around scientific content. He brings expertise in citizen science and public outreach.

Jeremy Kranowitz is a Senior Associate in The Keystone Center's Science and Public Policy Program. He manages an outreach and education program for the Department of Energy related to carbon sequestration, including education and outreach and risk communication training. Kranowitz will support sustainability and governance planning and implementation in DataNetONE.

John Kunze is a preservation technologist for the California Digital Library and has a background in computer science and mathematics. His recent work has focused on archiving websites, creating long-term durable digital references to information objects, and specifying lightweight (kernel) metadata, and includes collaboration on database preservation with the University of Edinburgh. With respect to DataNetONE, Kunze brings experience in data preservation as related to digital library systems.

Bertram Ludaescher is an Associate Professor in the Department of Computer Science at UC Davis. His research focus includes modeling, design, and optimization of scientific workflows and databases, data and workflow provenance, and knowledge representation and reasoning for scientific workflows. Ludaescher brings to DataNetONE expertise in the areas of data integration and semantics.

William Michener is Research Professor at the University of New Mexico's Department of Biology. He has experience in metadata management and the development of knowledge environments (e.g., Kepler scientific workflow system, EarthGrid). He brings significant science project management expertise, and engagement with two archives that are particularly relevant to the project—ESA's *Ecological Archives* and the LTER Network Information System.

Lorraine Normore is Assistant Professor in the School of Information Sciences at the University of Tennessee. She has developed significant, innovative user interfaces based on the principles of user-centered design. Her experience in gathering user input from working scientists and translating those data directly into design features will be invaluable for DataNetONE.

Ricardo Pereira is an independent Information Technology consultant based in Brazil, who has been researching and developing computer based tools for acquiring, managing, processing, and analyzing biodiversity information for nearly 10 years. He brings experience with the development, standardization, and deployment of metadata interoperability standards.

Line Pouchard is a research scientist at Oak Ridge National Laboratory's Computer and Computational Directorate specializing in metadata, ontologies, and interoperability issues of large scientific data. She has been involved in developing multi-institutional solutions for knowledge discovery and data access in climate science and other disciplines for over seven years. With regard to DataNetONE, she brings significant expertise as a contributor to the Department of Energy's Earth System Grid II that has been serving data and added-value services to (among others) the scientists of the Inter-governmental Panel on Climate Change who shared the Nobel Peace Prize in 2007.

Robert Sandusky is Assistant University Librarian for Information Technology and Clinical Associate Professor at the University of Illinois at Chicago's Richard J. Daley Library focusing on the long-term preservation of digital information. He brings to DataNetONE experience in designing, building, and operating highly-secure and reliable national-scale data communications networks and Web-based applications project management and development. His research focuses on human interaction with distributed socio-technical systems.

Ryan Scherle is the Digital Data Repository Architect at the National Evolutionary Synthesis Center. He leads the development of Dryad, a repository for evolutionary biology data. His research interests include distributed search systems and frameworks for interacting with diverse digital content.

Mark Servilla is the Lead Scientist of the Network Information System for the Long Term Ecological Research Network focusing on the design and implementation of a Network-wide information system to support the storage, discovery, and access to the LTER data holdings. He has 20 years of software development skills and familiarity of biological/ecological data management to ensure design principles of the DataNetONE architecture are broad-based and persist well into the future.

Kathleen Smith is Professor at Duke University and Director of National Evolutionary Synthesis Center. Her contribution to DataNetONE will be through leadership in data sharing initiatives in evolutionary biology as well as providing oversight for all NESCent activities associated with the project.

Carol Tenopir is a Professor in the University of Tennessee School of Information Sciences. Tenopir has studied patterns of scientific communication and scholarly publishing, in particular the use and design of digital publications for researchers and the role of the library with digital resources. She brings knowledge of the library and publishing communities, studying user needs and usage patterns of scholarly information, and working with and leading interdisciplinary research teams.

David Vieglais is a Senior Scientist at University of Kansas and was instrumental in the development of the Global Biodiversity Information Facility (GBIF). Vieglais brings 20 years of experience of practical design, development and deployment of a wide range of software and standards focused on environmental and ecological systems.

Von Welch is Co-Director for the National Center for Supercomputing Applications Directorate of CyberSecurity. He leads development of new security services and advanced CI and serves as PI for security projects with NSF, DoD and the FBI. He brings experience in federation security infrastructure for distributed projects such as DOEGrids and TeraGrid, and the Ocean Observatory Infrastructure.

Jake Weltzin is the Executive Director of the USA-National Phenology Network and an Associate Professor at University of Tennessee. He brings significant science project management experience, which he is currently applying to develop a continental-scale instrument for integrative assessment of global change that serves as an outreach and educational platform for citizens and educators.

Bruce Wilson is the Group Leader for Environmental Data Science and Systems at ORNL and the Manager for the NASA-funded ORNL Distributed Active Archive Center. He brings experience leading distributed cyberinfrastructure development teams supporting a broad range of science areas, a leadership role in three data archives, and 18 years experience working in a business environment.

Appendix A5. Role of Participating Organizations and Sectors

DataNetONE is comprised of numerous organizations with significant diversity, both in sectors represented and science expertise, through its PIs, Co-I's, Working Groups, and Node structure.

1. Role, Resources, and Capabilities DataNetONE PIs and Subawardees:

DataNetONE's lead organization, the **University of New Mexico (UNM)**, provides leadership for all aspects related to maturation of DataNetONE as well as the organization, development and implementation of one of the three DataNetONE Coordinating Nodes. UNM hosts the DataNetONE central office; employs the ED, ADs, and office and Coordinating Node staff and students; administers the subawards; coordinates Working Group, External Advisory Committee, and the DataNet International User Group meetings; and supports three of the ten Working Groups. UNM collaborates with the **New Mexico Computing Applications Center (NMCAC)** and **Intel Corporation** who will host the New Mexico Coordinating Node at Intel's Rio Rancho NM facility in conjunction with NMCAC's *Encanto*—presently, the third fastest supercomputer in the world. As cost-share, UNM provides office space (7 large offices that can be partitioned as necessary, one dedicated conference room, and receptionist and additional open space) in a secure facility (formerly the Joint Technology Office) at the university's Science and Technology Park, as well as access to three additional conference rooms located in the Park (to be used for Working Group meetings that occur in parallel). The UNM PI (William Michener) has extensive science management expertise, experience in developing metadata management solutions and knowledge environments (e.g., Kepler scientific workflow system, EarthGrid), and engagement with two archives that are particularly relevant to the project—the **Ecological Society of America's Ecological Archives** and the **Long Term Ecological Research (LTER) Network Information System (NIS)**. UNM also hosts one of the 26 LTER sites (Sevilleta LTER) and the LTER Network Office which coordinates activities across the 26 sites (approximately 1,500 scientists and students) and hosts the NIS for which Mark Servilla is the lead developer. Co-I Servilla brings his expertise with LTER NIS and the commercial sector to the DataNetONE Core CI Team.

The **National Center for Ecological Analysis and Synthesis (NCEAS)** at **UC Santa Barbara** hosts the second Coordinating Node and supports three of the ten Working Groups. Co-PI Stephanie Hampton initially co-leads the DataNetONE community engagement and outreach activities until the AD CE&O is hired and, thereafter, serves as a co-leader of the Community Engagement and Education Working Group co-leader. She also will oversee integration of Best Practices modules into existing informatics presentations to all NCEAS working groups, reaching hundreds of scientists and students during the first phase of DataNetONE. Co-I Matthew Jones serves on the Core CI Team and acts as co-leader of the research and development activities during year 1 until the AD D&O is hired. UC Santa Barbara and NCEAS, as part of their commitment to DataNetONE, provide the 24/7 operational environment for the Coordinating Node and high-speed bandwidth access, as well as furnished office space and conference facilities for the Node staff, students, and three Working Groups. NCEAS, through the leadership of Deputy Director Hampton and Director of Informatics Research and Development Jones, has extensive experience facilitating training, outreach, and interdisciplinary collaborative research by hundreds of scientists each year in its world class facilities. Furthermore, Co-I Jones provides leadership in earth observational science metadata standards, data management software, and scientific workflow systems.

Co-PI Kathleen Smith directs the **National Evolutionary Synthesis Center (NESCent)** at **Duke University** which supports three of the ten Working Groups. She will coordinate staff presentations of DataNetONE Best Practices to all NESCent working groups, reaching hundreds of scientists and students during the five year period. Co-I Ryan serves on the Core CI Team and provides expertise with respect to distributed search systems and frameworks for interacting with diverse digital content based on the experience in evolutionary biology. NESCent will provide the computing infrastructure and support personnel necessary to function as a Member Node, with primary data coming from Dryad, NESCent's repository for evolutionary biology data. NESCent will also provide office space for DataNetONE staff and conference facilities for three DataNetONE Working Groups.

Oak Ridge Campus hosts the third DataNetONE Coordinating Node. Co-PI Robert Cook serves as liaison with ORC's diverse facilities which include the Leadership Computing Facility, ORNL's **TeraGrid** node, and multiple data archive centers sponsored by **NASA**, **DOE**, and **USGS (ORNL Distributed Active Archive Center for Biogeochemical Dynamics, Carbon Dioxide Information and Analysis Center, and the World Data Center for Atmospheric Gases)**. ORNL Co-Is John Cobb and Line

Pouchard serve as Co-leaders of the Distributed Storage Working Group and the Data Integration and Semantics Working Group, respectively. ORC provides office space and 24/7 secure space for the Coordinating Node infrastructure. ORC personnel provide significant leadership with respect to centralized/distributed file management; community engagement and education; the development of standards, and metadata schemas, and distributed cyber infrastructure development. ORNL has been selected to host a **National Ecological Observatory Network (NEON)** core site and supports several successful programs targeting minority education institutions (including HBCUs) that will be used to encourage participation of underrepresented minorities in DataNetONE efforts.

The **USGS National Biological Information Infrastructure**, represented by Co-PI Mike Frame, serves as one of the initial “test” Member Nodes (deploying DataNetONE software prior to broader implementation) and is the principal liaison with the **World Data Center for Ecology and Biodiversity**, the **Global Biodiversity Information Facility (GBIF)**, and the **Inter American Biodiversity Informatics Network**. Co-PI Frame serves as co-leader of the Usability and Assessment Working Group. Co-I Vivian Hutchison initially co-directs DataNetONE community engagement and outreach activities until the AD CE&O is hired and, thereafter, serves as a co-leader of the Community Engagement and Education Working Group co-leader. She also will present Best Practices lectures and lead workshops for NBII stakeholders, including the **Organization of Fish and Wildlife Information Managers** (which includes agency representatives from all States in the US). USGS facilities, including NIST-certified computer facilities, office space, and administrative staff will be made available to DataNetONE on an as needed basis. The USGS NBII will also facilitate deployment of DataNetONE standards, protocols, and tools to federal libraries (including **USGS, EPA**, and others); provide expertise in management and operation of a distributed data center like the NBII; contribute to federal, state, and international outreach activities; and officially participating on international standards development committees.

The **University of Tennessee (UT)**, also serves as one of the initial DataNetONE “test” Member Nodes. Co-I Bruce Wilson (also affiliated with ORNL) serves on the Core CI Team and acts as co-leader of the research and development activities during year 1 until the AD D&O is hired. Co-I Suzie Allard coordinates UT activities and along with UT Co-I Carol Tenopir co-leads the Sociocultural Issues Working Group. UT Co-I Lorraine Normore co-leads the Usability and Assessment Working Group. UT provides Member Node CI and furnished office space for the staff and students. UT has a close collaboration with ORNL and other federal, state, private sector, and local organizations through managing research activities in the **Center for Information and Communication Studies (CICS)**. UT contributes expertise in understanding how scientists and engineers use and communicate information, electronic publishing, and how communication processes influence data use and preservation. CICS also brings expertise in the areas of strategic planning, performing service evaluations, modeling of scientific and technical communication, and in evaluating other science domain needs of DataNetONE activities.

The **Keystone Center** hosts annual meetings of the Long Term Sustainability and Governance Working Group which is co-led by Co-I Jeremy Kranowitz. The Center provides conference space and logistical support for the meetings. Keystone, under Director Peter Adler, designed the NBII organizational and governance structure, and brings significant expertise in sustainability, public policy, and experience with outreach and education programs for federal, state, and local governments.

The four remaining sub-awardees receive funds to support active participation in the Core CI Team. In return, the four organizations provide office space for their staff and connections to important sectors. Moreover, they will serve as DataNetONE Member Nodes (i.e., early adopters).

- a) Co-I Jeffery Horsburgh of **Utah State University** has extensive experience in developing and deploying cyberinfrastructure for data management, visualization, and analysis of hydrologic data and associated information about watersheds as part of his efforts in developing the Hydrologic Information System (HIS) for the **Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI)**; an organization representing more than a hundred US universities). The CUAHSI-HIS provides an initial CI framework and test bed for the **WATERS** environmental observatory effort.
- b) Co-I Dave Viegla of the **University of Kansas** has developed technical infrastructure for integrating biodiversity information at the global level (i.e. DiGIR, Species Analyst). Kansas brings significant biodiversity modeling expertise, access to millions of specimens records, and leadership in GBIF and the **Natural Science Collections Alliance**.
- c) The **Library at the University of Illinois-Chicago (UIC)**, under the direction of Co-I Robert Sandusky, will serve as a test site for the library software stack, assist in the development of online and face-to-face outreach materials, and conduct online and local face-to-face instructional sessions

for scientists, students, and librarians. UIC will also provide server, technical, and editorial resources for an open-access online journal to support reporting of research and operational findings from DataNetONE and other DataNet partner projects.

- d) Co-I Paul Allen of the **Cornell University Lab of Ornithology** has expertise in designing CI for collecting, archiving, and presenting biodiversity observations over the Internet, focusing on the involvement of amateur users and citizen scientists. Co-I Steve Kelling serves as co-leader of the Citizen Science and Outreach Working Group.

2. DataNetONE Plans for Active Participation by Diverse Sectors:

DataNetONE's structure inherently encourages active participation from a diversity of sectors. Working Groups and the DIUG promote transparency, user-centered design, and input from all stakeholders. In addition to those affiliations previously mentioned as sub-awardees, Co-leaders and pre-identified members of Working Groups represent many diverse sectors, including: (1) international and US academic institutions (**University of Edinburgh, University of Southampton, University of Manchester, University of Southern California, UC Davis, University of Michigan**); (2) high performance computing (**e-Science, National Center for Supercomputing Applications**); (3) governmental infrastructure projects (**Atlas of Living Australia**); (4) the library sciences (**California Digital Library, University of Illinois-Chicago, University of Tennessee, USGS**); (5) citizen science research networks (**USGS National Phenology Network, eBird, Project FeederWatch**); and professional societies (**Ecological Society of America**). Each of these institutions has far-reaching influence. For example, the California Digital Library: is associated with libraries and museums of ten University of California campuses engaged in collaboratively developing strategies for digital data preservation; has working relationships with the **California State Library** and **California public libraries**; is a member of **NISO** and the IIPC (a consortium of national libraries working together on web harvesting standards and practices); and partners with the **Library of Congress** in developing a national digital preservation strategy.

Working Groups will be further populated through an open recruiting process to assure broad sector representation and sufficient expertise in each area. As Working Group membership matures, the number of US and international institutions actively participating in the project will grow. We anticipate significant participation in Working Groups by computer scientists, library and information scientists, and informatics and domain experts associated with national and international environmental observatory networks (e.g., **NEON, WATERS, the Ocean Observatory Initiative, and the South African Environmental Observatory Network**).

The DIUG also facilitates broad and active participation by Earth observation data authors, users, and other diverse stakeholders including students, educators, researchers, libraries, data centers, professional societies, policy-makers, and the general public. Potential participants will be invited to become members of the DIUG through a variety of channels (i.e. announcements in journals, at professional meetings, and on listservs). Members of other DataNet networks will also be invited to the DIUG to foster dialog about design and development, and to share suggestions and ideas that are pertinent to all DataNet Partners. DIUG members will participate in identifying technical challenges and opportunities in education, research, and policy which can be addressed in the development of DataNetONE data products, tools, and services.

DataNetONE will seek to increase participation by researchers, libraries, and other institutions by actively communicating and engaging with key organizations to create awareness of DataNetONE and to solicit members of Working Groups and the DIUG. Examples of such organizations include:

- **Council of Graduate Schools** which has 480 members in the U.S. and Canada and whose members account for 90% of all doctorates awarded in the US and 75% of all US masters degrees.
- **Association of Commonwealth Universities** which has more than 500 institutional members in 53 independent states.
- **Coalition for Networked Information** which encompasses more than 200 institutions including academic, publishing, network and telecommunications, information technology, and libraries.
- **Digital Library Federation** (37 members including the US National Archives and Records Administration).
- **Association of Research Libraries** (123 libraries at the leading research institutions in the U.S. and Canada).

Appendix A6. Results from Prior Research

R. Bonney, **P. Allen**, C. Cooper, A. Dhondt, J. Dickinson. (NSF DRL-0540185, "Project NestWatch", \$1,894,268, 3/01/06 – 2/28/10). This project engages citizens to collect scientific observations of bird breeding, and harnesses the interest of Internet users to annotate and interpret millions of digital images of bird breeding. This project presents opportunities to prototype new user interface methodologies, data models, quality control processes, and approaches to observation validation and informs the design of generic tools for constructing observation systems.

R. Butler, and others. (SDCI - NMI- NEW: Observatory Middleware Framework, NSF 0721617, \$499,025, 9/1/2007-8/31/2010). Developing a generalized Observatory Middleware Framework to integrate existing and proposed observatory management technologies, and reduce duplication of functionality across observatories. **R. Butler** was also PI on two NSF NMI awards. The most recent was "Disseminating and Supporting Middleware Infrastructure: Engaging and Expanding Scientific Grid Communities." (co-PIs: C. Kessleman, ISI, and M. Livny, Wisc.; Award number 0330670 for \$180,000). This grant supported middleware integration, packaging, testing, hardening, dissemination and support for the NMI GRIDS Center middleware software release. It also supported numerous community outreach activities.

J. Cobb, and others. (ACI-0338605, SCI-0525789, SCI-0504074 totaling \$9,939,181) Creating bridges between large science user facilities and national scale cyberinfrastructure including construction and operations of a new TeraGrid site at ORNL with a fully functioning small local TeraGrid compute node and a close and continuing relationship with the Spallation Neutron Source and other facilities.

P. Cruse is the Director of the Web-at-Risk project, an NDIIPP funded four and one half-year effort led by the California Digital Library (CDL) to develop tools that enable librarians and archivists to capture, curate, preserve, and provide access to web-based information (NDIIPP Web-at-Risk Grant \$2,400,000, 3/01/05 – 9/01/09, continuation funds \$975,000, 01/01/08 – 06/30/09). The goal of the project is to build a sustainable web-archiving service that will enable libraries to work collaboratively to capture and preserve web-published materials that support the research, teaching, and learning at UC. The project is a multi-institutional project involving over 50 personnel and 14 institutions. **J. Kunze** is leading the technical development activities of the project.

E. Deelman was Senior Personnel on several NSF-funded ITR efforts (ITR-0086044, Grids Physics Network, 09/00-08/05, \$1,552,919; and AST-0122449, National Virtual Observatory, 10/01-09/06, \$1,085,635). In GriPhyN, Dr. Deelman led the effort to develop workflow technologies to support large-scale science. The Pegasus software was developed as part of this effort and used by LIGO for gravitational-wave analyses. In NVO, Dr. Deelman worked with the Montage scientists to fully parallelize the application using the workflow paradigm and today Pegasus is used to generate science-grade mosaics of the sky. This work resulted in numerous publications in both computer science and astronomy.

D. De Roure, **C. Goble**, and others (University of Southampton and University of Manchester). myExperiment - A Virtual Research Environment for Collaboration and Sharing of Experiments. JISC Capital Programme Virtual Research Environments phase 2. £414,466 (3/1/07 – 3/31/09). myExperiment makes it really easy for researchers to contribute to a pool of scientific workflows, build communities and form relationships. It supports the individual scientist on their personal projects, forming a distributed community with scientists elsewhere - enabling them to share, reuse and repurpose workflows, reduce time-to-experiment, share expertise and avoid reinvention. See <http://www.myexperiment.org>. **D. De Roure**, W. Hall, P. Henderson (University of Southampton). OMII-UK Centre and Managed Programme. EPSRC Reference EP/D076617/1. £6,657,560 (7/1/06 – 12/31/10). OMII-UK enables advanced e-infrastructure solutions based on open source Grid middleware components which are engineered to a high quality, interoperable and easily-used, drawing on the software generated by the UK e-Science programme. See <http://omii.ac.uk>

C. Duke. (NSF DEB-0424702, "An Ecology, Evolution, and Organismal Biology Societies Summit Meeting: Critical Steps Toward a Biological Data Systems Confederation," \$80,000, 5/15/04 – 4/30/07). This project initiated a continuing effort by the Ecological Society of America (ESA) to promote the development of resources and the reduction of cultural obstacles to data sharing. **C. Duke**. (NSF DEB-0533052, "Joint Working Group on Data Sharing and Archiving: Continuing Steps Toward a Biological Data Systems Confederation," \$80,000, 4/15/06 – 3/31/09). This project continues the efforts to promote

data sharing begun at the Society Summit. Three follow-on workshops have explored questions about data registries, data centers, and cultural obstacles to data sharing, respectively. To date, 22 professional societies have participated in workshops, along with representatives of 26 other organizations.

M. Frame, V. Hutchison, and others participate in the “National Biological Information Infrastructure” (U.S. Geological Survey, \$12,000,000 / year, 10/1/2000 – ongoing). The network provides access, management, and tools related to biological data and information discovery, management, and visualization. Cyberinfrastructure supports collaboration, data replication, geospatial (NBII Geospatial Interoperability Framework), taxonomic (Integrated Taxonomic Information System), thesaurus (NBII Biocomplexity Thesaurus), standards (FGDC BDP, NBII Dublin Core+, etc.), and other network services.

M. Frame and individuals from over 40 international institutions are active in “Building the Inter-American Biodiversity Information Network (IABIN; World Bank, \$6,000,000, 6/29/04-6/30/10). The project is creating CI to enable search and access of biodiversity data and information within the Americas. **M. Frame** is also participating in the development of the World Data Center for Biodiversity and Ecology. The World Data Center incorporates data of appropriate ICSU scientific programs or monitoring activities in the areas of biodiversity and ecology. Free access to data, information products, and related services are available to scientists and interested users in any country.

C. Goble and others. (Universities of Manchester, Newcastle, Nottingham and Southampton). myGrid - A Platform for e-Biology. EPSRC Reference EP/C536444/1. £419,572 (7/11/05 – 1/10/09). The myGrid consortium has (a) successfully pioneered world-class research on the Semantic and Data Grid, and (b) developed open source high-level service-based middleware to support in silico experiments in biology that are in use by biologists and being adopted by a wide range of national projects and international partners. See <http://mygrid.org.uk>. **C. Goble** and others. (University of Manchester). myGrid: An OMII-UK Node - Services and Middleware for e-Science. EPSRC Reference: EP/D044324/1 £1,944,847 (3/1/06 – 8/31/09). myGrid is one of the long-established projects that has produced popular and useful middleware for Life Scientists such as the workflow workbench Taverna. See <http://mygrid.org.uk>. **C. Goble** and others. (University of Manchester). ESNW : A Centre for Collaborative Multidisciplinary e-Research in the North West. EPSRC Reference EP/D057248/1. £232,149 (3/1/07 – 2/28/10). ESNW was established to provide support for Grid and e-Science activities in the region and to provide effort in building a UK National Grid. See <http://www.esnw.ac.uk>

D. Hobern, S.Blum (California Academy of Science), **R. Pereira** (Senior Personnel). (Gordon and Betty Moore Foundation, Grant Award Letter Agreement 439, "Hardening TDWG Data Standards Development Process", USD 1,494,000, 7/01/05 – 12/31/07). This project is modernizing the Taxonomic Databases Working Group as an international body for development of biodiversity information standards, including developing an integrated standards architecture and developing tools and documentation around existing standards. TDWG standards relate particularly to scientific names and classification, biological specimens, field observations, organism descriptions and identification keys - see <http://www.tdwg.org/>.

P. Honeyman, W.A. Adamson, and J.B. Fields. (“SciDAC: Petascale Data Storage Institute”, U.S. Department of Energy, Office of Science, Cooperative Agreement No. DE-FC02-06ER25766, \$1,498,405, September 2006–August 2011). The Petascale Data Storage Institute focuses on the data storage problems found in petascale scientific computing environments, with special attention to interoperability, community buy-in, and shared tools. **P. Honeyman**, W.A. Adamson, and S. McKee. (“NMI: GridNFS,” National Science Foundation, SCI-0438298, \$1,254,981, September 2004–August 2008). GridNFS extends distributed file system technology and flexible identity management techniques to meet the needs of grid-based virtual organizations.

J. Horsburgh participates in NSF EAR 0622374 (“GeoInformatics: Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) Hydrologic Information Systems,” \$4,500,000, 1/15/07 – 1/15/12. This five-year project is focused on development of Hydrologic Information Systems for advancing Hydrologic Science and development of cyberinfrastructure for supporting Hydrologic and environmental observatories. The CUAHSI Hydrologic Information System (HIS) is a geographically distributed network of hydrologic data sources and functions that are integrated using web services so that they function as a connected whole. The HIS integrates national water data archives with locally published hydrologic data, and makes them directly accessible to hydrologic scientists. **J. Horsburgh** participates in NSF CBET 0610075 (“Tools for Environmental Observatory Design and Implementation: Sensor Networks, Dynamic Bayesian Nutrient Flux Modeling, and Cyberinfrastructure Advancement,”

\$355,936, 11/1/06 – 10/31/08. This project is one of several WATer and Environmental Research Systems (WATERS) Network "test-bed" projects funded to evaluate issues that need to be resolved for cost-effective design of environmental observatories for research on interactions among hydrological, physical, chemical, and biological processes at the watershed scale.

V. Hutchison. (Federal Geographic Data Committee (FGDC), Award numbers: IA760310268 (2007); IA660110253 (2006); IA560110409 (2005); IA460110490 (2004); 2003; 2002. This FGDC award category provides funding to organizations to assist in the implementation of FGDC-endorsed standards. The NBII projects identify metadata needs of NBII partners (including USGS Science Centers), assist in the development of a sustainable metadata program for the organizations by providing necessary training in the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) and the Biological Data Profile (BDP), and provide support for record discovery through the NBII Clearinghouse.

M. Jones and others from 6 institutions in the US. (NSF 0619060, "Management and Analysis of Environmental Observatory Data using the Kepler Scientific Workflow System", \$2,700,112, 10/1/06 – 9/30/10). This four-year project is combining the Kepler scientific workflow system with dynamic real-time data grids associated with multiple sensor networks. This near Real-time Environment for Analytical Processing (REAP) will provide an open-source, extensible and customizable framework for designing and executing scientific models that consume data streams from archives and sensor networks. REAP is working to unify the programmatic interfaces for accessing data from diverse systems, including data archives such as the KNB that use the EarthGrid protocol, oceanographic data servers that use the OPeNDAP protocol, and sensor networks that use vendor-specific proprietary protocols.

S. Kelling, P. Allen, R. Bonney. (NSF DUE-0734857, "The Biodiversity Analysis Pipeline", \$497,420, 1/1/08 - 1/1/10). The Biodiversity Analysis Pipeline will allow anyone to access, organize, and visualize the primary biodiversity data currently available online via scientific analyses and workflow applications through an intuitive and rich web interface. Users will be able to create new visualizations, or repurpose existing visualizations, which are unique and valuable for understanding ecological systems. **S. Kelling** and others. (NSF IIS-0612031, "SEI+II Ecological Discovery and Inference: Tools for Data-driven Exploration and Testing of Observational Data", \$987,000, 8/1/06 - 8/1/09. Through a new synergy between high-performance computing, statistics, computational biology, and computer science we will organize, analyze and disseminate this unique assemblage of natural history data and advance conservation and management of biodiversity at many ecological scales. This project will create new techniques for interactive exploration and analysis of massive spatio-temporal data collections, to generate statistical descriptions and tests of long-term population trends, geospatial patterns of distribution, and seasonal movements of bird populations. **S. Kelling**, D. Fink, G. Ballard, A. Dhondt. (NSF DBI- "Multi-Scaled Data in Ecology: Scale Dependent Patterns in the Environment", \$770,000, 8/15/06 - 8/15/09. The goal of this project is to expand the resources and analysis tools available at the Avian Knowledge Network (AKN) (<http://www.avianknowledge.net> <<http://www.avianknowledge.net/>>), by developing a new analytical methodology, Hierarchical Semiparametric Modeling (HSM), which combines the power of non-parametric data mining strategies, with parametric statistics.

P. Adler, D. Thompson, **J. Kranowitz** and others. (U.S. Geological Survey Informatic Node Strategic Planning, 2005 – 2006, \$222,705). Led two diverse workgroups (Northeast Information Node and Invasive Species Information Node) to develop strategic plans on how best to collect, manage, and disseminate information on biological taxa that is collected by the USGS and other institutions. Through conference calls and strategic planning meetings, helped guide diverse groups to develop a strategic plan, including gap analysis, establishing screening criteria, and writing draft consensus documents.

B. Ludaescher is currently the PI of NSF/SDCI "Kepler/CORE" (OCI-0722079), *A Comprehensive, Open, Robust, and Extensible Scientific Workflow Infrastructure* (\$1,700,000); NSF/BIO+II "Kepler/Chip-chip" (IIS-0612326), *A Collaborative Scientific Workflow Environment for Accelerating Genome-Scale Biological Research* (\$600,139), and a co-PI of "AToL/pPOD" (IIS-0630033), a collaborative project with U Penn, Yale, and U Florida on *Core Database Technologies to Enable the Integration of AToL information* (\$462,000), among others. In his projects Dr. Ludaescher and his group at UC Davis conduct R&D for user-oriented scientific workflow design, workflow reuse, optimization, and archival. These approaches are informed by static type inference, other semantic enhancements, and provenance information. His computer science research in these areas is driven by the practical challenges arising from the collaboration with biologists, ecologists, and physicists.

D. Williams, D. Bernholdt, **L. Pouchard**, and others from 5 institutions in the US. (DOE Office of Science, "Earth Systems Grid II (ESG)," 2000-2005, \$8.7 M, and "Scaling the Earth System Grid to Petascale Data Center for Enabling Technology," starting 2006, approximately \$2.75 million per year for five years). ESG provides a seamless and powerful environment that enables the next generation of climate research through a combination of Grid technologies and emerging community technology, distributed federations of supercomputers and large-scale data and analysis servers. ESG currently holds and serves 300 Terabytes of annotated data to over 8000 users worldwide and growth to Petabytes is expected by 2010. These holdings include 35 TB provided to IPCC scientists who jointly won the 2007 Nobel Peace Prize.

C. Tenopir. (Institute of Museum and Library Services LG-02-04-0034-04, "Maximizing Library Investments in Digital collections Through Better Data Gathering and Analysis (MaxData)", \$446,988, 12/31/04-9/30/08). The purpose of this project is to provide librarians with cost and benefit information about the main methods of gathering and analyzing usage data for electronic collections (including: COUNTER-compliant reports supplied by vendors; those that are not COUNTER-compliant; locally gathered log data, including deep log analysis; and surveys.). **C. Tenopir, R. Sandusky.** (CSA (ProQuest), "Value of CSA Deep Indexing for Researchers", \$74,909, 5/1/06-5/31/07). A multiple data collection methods approach was used to provide both quantitative and qualitative analysis at varying levels to study the utility and usefulness for academic and other researchers of CSA's newly developed feature ("deep indexing") that allows indexing of figures and tables in journal articles in CSA databases (patent pending). **C. Tenopir, S. Allard, K. Levine.** (Institute of Electrical and Electronics Engineers, Inc. 05-1440-06, "How Technical Professionals Work", \$79,263, 3/8/05 -12/31/06). The project was designed for this professional organization to study how design technical professionals in high tech industries communicate and use information in their work. **C. Tenopir, S. Allard, K. Levine.** (Institute of Electrical and Electronics Engineers, Inc. 06-3252-07, "How Technical Professionals Work – Part 2--China", \$24,270, 9/1/06-9/30/08). Project is a continuation of a previous study of how design technical professionals in high tech industries communicate and use information in their work.

D. Vieglais has been PI, Co-PI and Senior Personnel on several NSF-Funded projects: (Distributed Information Network for Avian Data, DBI-9808739, 9/25/98 – 11/5/02; \$600,083; Development of an Integrated Network for Distributed Databases of Mammal Specimen Data, DBI-0108161, 9/13/01 – 8/31/06, \$1,642,754; The HerpNet Community Informatics Project: Development of a Distributed Information Network of North American Herpetological Databases (HerpNet), DBI-0132303, 9/15/02 – 8/31/08, \$2,522,253; ORNIS: A Community Effort to Build an Integrated, Distributed, Enriched, and Error-checked ORNithological Information System, DBI-0345448, 7/10/04 – 7/31/08, \$1,391,989; Collaborative Research: Building the Information Community Infrastructure - A Test Case Implementation for Ichthyological Collections, DBI-0415600, 7/15/04 – 7/31/08, \$741,760). Dr. Vieglais led software design and development for the information sharing components of these projects which have been instrumental in providing a global infrastructure for sharing biodiversity information.

V. Welch led "Policy Controlled Attribute Framework" (NSF SCI/NMI award number 0438424, \$396,060, 12/1/04-11/30/07, collaborative with award number 0438385 Katarzyna Keahey, U Chicago) to develop software that allowed for the interoperability between public key infrastructure computational grids. Welch was CO-PI on the NCSA subaward for the Ocean Observatories Initiative (JOI/NSF OCR-0418957, \$53,690, 9/1/08-2/29/08), responsible for designing the identity management architecture for federation of the three observatories (global deep sea, coastal, and regional cabled) comprising the OOI.

J. Weltzin with two other universities. (NSF-DEB-0418363, 2004-2008, \$72,651, Collaborative Research: Vulnerability of Semi-arid Grasslands to Encroachment by Woody Plants: the Role of Grass Invasions, Seasonal Precipitation, and Soil Type). A large-scale manipulative experiment designed to determine impacts of climate change, grass invasions, and woody plant dynamics on semi-arid savannas. Project supported 5 graduate students including four women, one Native American, and one Hispanic, and many undergraduates, including minority women, and has produced 10 papers to date. **J. Weltzin** with Oak Ridge Nat'l Laboratory. (DOE-OBER-PER DE-FG01-05ER05-02 TN, 2001-2007, \$1,061,653, "Community and ecosystem response to global change: interactive effects of atmospheric carbon dioxide, surface temperatures, and soil moisture"). A large-scale manipulative experiment designed to determine impacts of elevated atmospheric [CO₂] concentration, increased air temperature, and changes in soil moisture using open-top chambers containing constructed ecosystem with plants typical of an old-field plant community.

Letters of Collaboration from Institutions Integral to DataNetONE

Atlas of Living Australia – Donald Hobern (Director)
amazon.com (Amazon Web Services) – Adam Selipsky (Vice President, amazon.com)
Cornell Laboratory of Ornithology – Steve Kelling (Director, Information Science Program)
Ecological Society of America – Clifford S. Duke (Director of Science Programs)
Global Biodiversity Information Facility – Meredith A. Lane (Public & Scientific Liaison)
Innovation Valley Partners (Battelle Ventures Affiliate; associated with Oak Ridge National Laboratory and the DOE National Laboratory system) – Glenn Kline (General Partners)
Keystone Center – Jeremy Kranowitz (Senior Associate)
National Evolutionary Synthesis Center – Kathleen K. Smith (Director)
New Mexico Computing Applications Center and Intel Corporation – William J. Feiereisen (Interim Science Director)
Oak Ridge National Laboratory, Environmental Data Science and Systems – Bruce E. Wilson (Manager, ORNL NASA Distributed Active Archive Center)
South African National Parks and the South African Environmental Observatory Network – Judith Kruger (Program Integrator)
Taiwan Forestry Research Institute/Taiwan Ecological research Network – Chau Chin Lin (Senior Scientist / Coordinator of Information Management)
United States Department of Interior (United States Geological Survey; National Biological Information Infrastructure; Committee on Environment and Natural Resources; Science.gov; Organization of Fish & Wildlife Information Managers; World Data Center for Biodiversity and Ecology; Inter-American Biodiversity Information Network) – Gladys A. Cotter (U.S. Geological Survey, Associate Chief Biologist for Information)
University of California, California Digital Library – Laine Farley (Interim Executive Director)
University of California, Davis – Bertram Ludaescher (Associate Professor and Director of the Data and Knowledge Systems Lab)
University of California, Santa Barbara – William Murdoch (Interim Director, NCEAS), Mark Brzezinski (Director, Marine Science Institute), Brenda Johnson (University Librarian)
University of Illinois at Urbana-Champaign – Thom Dunning (Director, National Center for Supercomputing Applications)
University of Illinois at Chicago, Richard J. Daley Library – Mary M. Case (University Librarian)
University of Kansas, Biodiversity Research Center – Leonard Krishtalka, Director
University of Manchester (UK) – Professor Carole Goble (Director, ^{my}Grid Project and ESNW)
University of New Mexico – John K. McIver (Interim Vice President for Research and Economic Development)
University of Southampton (UK) – David De Roure (Professor of Computer Science)
University of Southern California, Information Sciences Institute – Ewa Deelman (Project Leader, Pegasus Workflow Management System)
University of Tennessee, College of Communication and Information – Michael O. Wirth (Dean)
USA National Phenology Network – Jake Weltzin (Executive Director)
Utah State University – Norma Buxton (Sponsored Programs Administrator)

Dr. William Michener
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
Department of Biology, MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

13 March 2008

Dear Dr Michener,

The *Atlas of Living Australia* is a five-year project funded by the Australian Government's NCRIS programme to "develop an authoritative, freely accessible, distributed and federated biodiversity data management system that links Australia's biological knowledge with its scientific reference collections and other custodians of biological information".

The project aims:

- To integrate data on specimens held by Australia's natural history collections and data from field observations of living organisms
- To support the management and integration of biological data from all areas of research (molecular to ecological)
- To develop search interfaces and web services to facilitate discovery of biological information resources and to support the use of biological data in scientific research, policy-making and education
- To ensure that data relating to Australian organisms is well-managed and organised to meet future information requirements.

As part of this work, the *Atlas of Living Australia*, will be working with other NCRIS projects to develop common approaches to management of data and metadata, use of access frameworks and supercomputing infrastructure, and long-term management of data for use in science and policy development.

For these reasons, we believe that the DataNetONE initiative is a significant and highly relevant activity for the *Atlas of Living Australia* as it seeks to meet its own goals. We look forward to the opportunity to collaborate with you in establishing standards and policies and in working to promote full interoperability of data resources at a global scale.

As both projects develop, the *Atlas of Living Australia* will seek to establish a DataNetONE Member Node and to explore synergies in software development and in the use of biodiversity data.

We wish you success in your proposal.

Best wishes,



Donald Hobern
Director, Atlas of Living Australia
Donald.Hobern@csiro.au
+61 2 6246 4352



March 6, 2008

Dr. William Michener
Long-Term Ecological Research Network Office, Associate Director
Department of Biology, MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dear Bill,

Amazon Web Services is excited to collaborate with you in your DataNetONE proposal to the National Science Foundation. Amazon Web Services provides developers direct access to the robust technological building blocks that support Amazon. The following sets forth the terms of a proposed arrangement between Amazon Web Services and DataNetONE initiative regarding the provision of certain services by Amazon in support of the initiative.

First, we may provide detailed technical information to your Core CI Team on the functionality and characteristics of Amazon Simple Storage Service (S3) and the Elastic Compute Cloud (EC2), as well as other web services as appropriate. We will also work closely with the Core CI Team to devise integration approaches to our services, and help fit appropriate Amazon web services into your overall architecture.

Second, we may work with your Long Term Governance and Sustainability Working Group to model the costing out for systems built on Amazon web services. We anticipate participating directly in those Working Group meetings at the Keystone Center in Colorado, but will have personnel available to remotely interact with the Working Group in the event of schedule conflicts.

Third, many of the initial research questions identified for several of the other nine DataNetONE Working Groups are of direct relevance to Amazon Web Service development efforts. We may work with you in identifying opportunities for additional collaboration in these Working Group activities. Such collaboration might entail participation in one or more Working Groups or sharing technological expertise and solutions. We can pass along best practices for DataNetONE that we have garnered from working with other corporate and academic enterprises that are building similar systems.

Fourth, we may bring awareness to DataNetONE and those components that are based on our technological solutions to the extent that this is deemed as mutually beneficial.

Finally, I have enjoyed learning more about DataNetONE and brainstorming with you about the opportunities for a real and meaningful collaboration between our corporate research and development team and the multi-disciplinary, international team that comprises DataNetONE.

This letter has been prepared solely to identify the preliminary scope of our relationship and significant matters needing further discussions with a goal of reaching one or more definitive agreement(s). Therefore, nothing herein shall be interpreted as a binding offer or agreement as to any terms or conditions.

Please let me know how we can continue to contribute to your proposed effort. I look forward to hearing back from you on the next steps and to working with you and your colleagues.

Regards,

A handwritten signature in black ink, appearing to read "Adam Delipish". The signature is fluid and cursive, with the first name "Adam" and last name "Delipish" clearly distinguishable.



13 March 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The mission of the Cornell Lab of Ornithology (CLO) is to interpret and conserve the earth's biological diversity through research, education, and citizen science focused on birds. That mission has led CLO to compile one of the world's largest datasets of biological observations, numbering over 30 million records from across the Western Hemisphere. CLO also heads the Avian Knowledge Network which has federated an additional 11 million records from partners throughout the US and Canada. We are constantly seeking new means to share, integrate, and analyze these data to advance science and conservation.

We are at a critical junction in our need for a comprehensive data management strategy in Ecology. Now we must go beyond simple data curation and provide the opportunity for data synthesis, exploration, and new analysis. The DataNetONE (Observation Network for Earth) goal to build an infrastructure enabling management and long term preservation and access to scientific datasets is functionality critical to making the best use of datasets of all sizes covering a wide range of scales and domains held by CLO and other organizations.

CLO agrees to collaborate as a sub-awardee in the DataNetONE project. As part of that collaboration, I will serve as co-leader of the DataNetONE Citizen Science and Outreach Working Group. In addition, Paul Allen, one of CLO's lead software system architects, will serve as a member of the DataNetONE Core Cyberinfrastructure Team. CLO also commits equipment, space, and systems support for a DataNetONE member node beginning in the project's first year. We will serve as a test site for the DataNetONE service interface and investigator toolkit. We look forward to sharing our large observation datasets and using other datasets in the DataNetONE network to enrich the analyses and data mining of the avian observations we hold.

Sincerely,

Steve Kelling
Director, Information Science Program
Cornell Lab of Ornithology



The Ecological Society of America
1990 M Street NW
Suite 700
Washington, DC 20036

7 March 2008

Dr. William Michener
Department of Biology
University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The Ecological Society of America (ESA) is pleased to collaborate with you and your team on your DataNetONE proposal to the National Science Foundation's DataNet competition. ESA's 2004 "Visions report," *Ecological Science and Sustainability for a Crowded Planet*, recommended that the Society "promote the standardization of data collection, data documentation, and data sharing." Since that time, ESA's Office of Science Programs has followed this recommendation by hosting a series of NSF-sponsored workshops to formulate common data sharing policies among professional societies in ecology, evolution, and organismal biology, and to develop recommendations for new resources and training for data sharing. We are grateful for the participation by you and other DataNetONE team members in these workshops and for your leadership of related ESA data sharing and preservation efforts, particularly your role as Editor of our *Ecological Archives*.

The goals and methods of the proposed "DataNet Observation Network for Earth (DataNetONE)" are fully consistent with the recommendations of the Visions report and these workshops. ESA therefore welcomes the opportunity to participate as a member of DataNetONE.

Sincerely,

A handwritten signature in dark ink, reading 'Clifford S. Duke'. The signature is fluid and cursive, with the first name 'Clifford' being more prominent.

Clifford S. Duke, Ph.D.
Director of Science Programs



Mr. Mike Frame
USGS Center for Biological Informatics
Building 1916T2
230 Warehouse Road
PO Box 6015
Oak Ridge, TN 37381, USA

4 March 2008

Dear Mike:

Thank you so much for telling GBIF about the DataNet proposal.

GBIF works in two major areas: 1) development of standards, protocols and other elements of the biodiversity information architecture and 2) promoting and encouraging increase in quality biodiversity data content, both of names and occurrences, via our data portal (data.gbif.org).

Of course, this work is highly dependent on the data served and shared with other initiatives to be perpetuated through time. It would be disastrous for GBIF, and indeed for all initiatives and people everywhere who are coming to depend on the persistence of data and services on the Internet.

Thus, GBIF enthusiastically welcomes the advent of a project designed to enable the long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data. This is especially true as GBIF is itself focusing this year on improving our capacities for handling observational data.

GBIF looks forward to collaborating in the development of community-derived interoperability standards and the incorporation of new value-added and innovative technologies into the grand process in which we are all engaged - which is nothing less than the reinvention of the concepts of "library", "encyclopedia" and "data store" as a new way of sharing information among peoples, across both space and time.

Sincerely yours,

Meredith A. Lane
Public & Scientific Liaison

**Global Biodiversity
Information Facility
Secretariat**

Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark

Tel.: +45 35 32 14 84
Fax: +45 35 32 14 80
Email: mlane@gbif.org
Web: www.gbif.org

March 14, 2008

Dr. William K. Michener
Long-Term Ecological Research Network Office, Associate Director
Department of Biology
MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dr. Michener:

Innovation Valley Partners is pleased to collaborate with the DataNetONE project team, particularly in the area of ongoing financial sustainability. Innovation Valley Partners is the \$35 million affiliate fund of Battelle Ventures, L.P., a \$220 million independent fund that invests in early-stage companies nationwide. Backed by business leaders in Knoxville, Tenn., Innovation Valley Partners participates alongside Battelle Ventures, focusing on companies in the areas of health & life sciences, security, and energy & environment. Because of its relationship with Battelle Ventures, Oak Ridge National Laboratory, and the DOE National Laboratory system, Innovation Valley Partners is particularly experienced in the issues of appropriate and sustainable commercialization and monetization of technologies generated as the result of publicly funded research.

While we understand that the software tools developed by DataNetONE will be licensed as Open Source and the data holdings of DataNetONE will be publicly accessible, there are numerous examples of cases where open source communities are able to create subsidiary entities which generate a revenue stream for at least partial support of ongoing development and innovation. A common example is the creation of a services firm which can assist for-profit corporations in the use and customization of the open source technology for the specific needs of that corporation. In the case of DataNetONE, such an arrangement would enable the for-profit corporation to better manage its own data, ensure that innovations in the DataNetONE software arising from such deployments flow back into the core tool sets, and provide a revenue stream to contribute to the ongoing sustainability of DataNetONE.

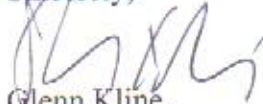
To help enable the long-term sustainability of DataNetONE, Innovation Valley Partners proposes to collaborate with DataNetONE in the following ways:

- Having Glenn Kline serve on the Governance and Sustainability Working Group

- Enable the participation of DataNetONE (and potentially other DataNet partner teams) in our entrepreneurial forums, which will allow DataNetONE connect and exchange ideas with start-up companies and other venture capitalists. This particularly includes involving DataNetONE with our entrepreneurs in residence program, which includes some entrepreneurs from software companies and consulting firms.
- Provide general advice on issues of long-term sustainability, business models, and incorporation strategies. While DataNetONE would need to secure its own legal counsel to proceed on any incorporation and the creation of any subsidiary entities, Innovation Valley Partners and the entrepreneurs with whom we work do have relevant experience which can provide useful information to help guide DataNetONE in the many options for these types of decisions.

We very much look forward to a productive and interesting collaboration.

Sincerely,



Glenn Kline,
General Partner

March 20, 2008

Dr. William K. Michener
Long-Term Ecological Research Network Office, Associate Director
Department of Biology
MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dr. Michener:

The Keystone Center (TKC) agrees to collaborate as a sub-awardee in the University of New Mexico's DataNetONE (Observation Network for Earth) project. As a project collaborator, TKC will provide

- Conference room space and associated equipment, including projectors, screens, and meeting materials to support the sustainability work group for the first five years of the project.
- Assist, as needed, in the development of online and face-to-face outreach and educational materials for scientists, students, and citizen scientists.
- Mr. Jeremy Kranowitz will serve as a member of the Sustainability Team.

We look forward to a rewarding and productive research effort.



Jeremy Kranowitz
Senior Associate
The Keystone Center



National Evolutionary Synthesis Center

2024 W. Main St., Suite A200
Durham, NC 27705 USA
<http://www.nescent.org>
919-668-4551

March 12, 2008

Dr. William K. Michener
Long-Term Ecological Research Network Office
Department of Biology
MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dr. Michener:

The National Evolutionary Synthesis Center (NESCent) is proud to participate in the development of the DataNetONE network. Archiving and preserving data are core to NESCent's mission, and the DataNetONE architecture will complement NESCent's repository of evolutionary biology data, Dryad.

NESCent is jointly operated by Duke University, the University of North Carolina at Chapel Hill and North Carolina State University. It receives its core funding from the NSF. NESCent sponsors synthetic research in evolutionary biology among scientists throughout the world, helps to build the informatics tools needed for such synthetic research to take place, and also has a strong education and outreach program. Among NESCent's major informatics goals is the preservation, sharing, and synthesis of published data in evolutionary biology. In collaboration with a set of major evolutionary biology journals, we are developing a digital repository named Dryad that aims to capture complete data packages, with associated metadata, from authors at the time of publication.

As a sub-awardee, NESCent will:

- provide equipment, space, and systems support for a DataNetONE member node, with primary data coming from Dryad.
- facilitate relationships with journals and societies in the evolutionary biology community.
- provide office space for a DataNetONE postdoc.
- facilitate interactions between the DataNetONE postdoc and the hundreds of scientists that visit NESCent each year.
- assist in the development of online and face-to-face outreach and instructional materials for scientists, students, and citizen scientists.
- host all meetings for three of the DataNetONE working groups, including logistics support, meeting space, and technical support.

In addition, NESCent staff member Ryan Scherle, architect for the Dryad repository, will serve as a member of the Core Cyberinfrastructure Team.

We believe that the full value of the contents in Dryad will only be realized when researchers and students can easily relate those data to the wide variety of biodiversity, ecological, and environmental data available from other centers through a virtual network, which is why we place great importance on the current proposal. The participants in your project are leaders in the development of standards, technologies, and capabilities for data sharing in biodiversity, ecology, environmental science, and library science, and so we thus have full confidence in this effort. We look forward to working with you more closely once it is funded.

Sincerely,

A handwritten signature in black ink, appearing to read 'Kathleen K. Smith'.

Kathleen K. Smith
Director, National Evolutionary Synthesis Center
Professor of Biology, Duke University

Dr. William Michener
Long Term Ecological Research Network Office
University of New Mexico
MSC03 2020
Albuquerque, NM 87131-0001

3 March 2008

Dear Bill,



The New Mexico Computing Applications Center (NMCAC) and Intel Corporation are pleased to collaborate with you and the exceptional team that has been assembled as part of your DataNetONE proposal to the National Science Foundation's DataNet competition. Under the leadership of Governor Bill Richardson, NMCAC was created to leverage advanced supercomputing and storage resources with research and development projects associated with New Mexico's universities, industry, and national laboratories (Sandia and Los Alamos). The overarching goal of the Center is to enable scientists to solve the complex problems and challenges that affect the planet and the communities of our state and nation.

The centerpiece of NMCAC is *Encanto*—a 14,336-core SGI(R) Altix(R) ICE supercomputer system from SGI that is outfitted with Intel(R) Xeon(R) processors, 28 Terabytes (TB) of memory, and a 172TB SGI(R) InfiniteStorage 4500 solution. The integrated high performance computer (HPC) system is housed at the Intel Corporation facility in Rio Rancho, N.M. and is accessible via the National LambdaRail and Internet2. Although the HPC landscape is constantly changing, *Encanto* currently ranks as the third fastest supercomputer in the world. Moreover, it is anticipated to drive cutting edge research in the state for the next several years.

I and several of my colleagues in the Governor's office and at the New Mexico Computing Application Center are especially pleased to see that DataNetONE is initially focused on preserving and sharing Earth observation data from an array of biological, geophysical, and environmental sciences. We expect that some of *Encanto*'s computing cycles will be used in fine-scale weather and climate forecasting, as well as solving the complex challenges associated with better understanding the linkages among climate, water use, and human activities. Because the goals of NMCAC and DataNetONE are so closely aligned, NMCAC and Intel Corporation have agreed to co-locate DataNetONE's cyberinfrastructure (servers and 35+ TB storage) at Intel's Rio Rancho facility, providing space for the racks as well as 24/7 power, security, and high-bandwidth access.

We very much look forward to working with you and your international team throughout the lifetime of DataNetONE. In addition to hosting DataNetONE cyberinfrastructure, we are enthusiastic about working with you and your partnering institutions to ensure that both DataNetONE and NMCAC are sustained for decades into the future so that we may collaboratively address some of our Nation's most pressing scientific challenges.

Sincerely,

A handwritten signature in black ink, appearing to read "Wm Feiereisen". The signature is fluid and cursive, with the first name "Wm" and last name "Feiereisen" clearly distinguishable.

Dr. William J. Feiereisen
Interim Science Director
New Mexico Computing Applications Center
Governor Bill Richardson's Office
Suite 400, 490 Old Santa Fe Trail
Santa Fe, NM 87501

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Bruce E. Wilson
P.O. Box 2008, MS 6407
Oak Ridge, TN 37831-6407
Phone: (865) 574-6651
Fax: (865) 574-4665
wilsonbe@ornl.gov

March 10, 2008

Dr. William Michener
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
Department of Biology, MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Bill,

The Environmental Data Science and Systems Group at Oak Ridge National Laboratory is pleased to be a collaborator for the DataNetONE proposal. A separate letter from the Department of Energy will provide formal authorization for Bob Cook to participate as a co-PI, John Cobb to participate as a working group leader, and Line Pouchard to participate as a working group leader. That letter will also authorize ORNL's participation as a coordinating node, in collaboration with the University of Tennessee, and the provision of appropriate physical hosting facilities with qualified systems administration. To help ensure the success of the collaboration between ORNL and UT and as a cost-saving measure to DataNetONE, I will participate through a joint appointment with the University of Tennessee in the School of Information Sciences. While some details remain to be sorted out, this joint appointment will most likely involve the ORNL-UT Joint Institute for Biological Sciences (JIBS), which opens up further potential for broad biological sciences participation in DataNetONE.

The Environmental Data Science and Systems (EDSS) Group includes three substantial ecological data centers: the NASA-funded ORNL Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC), the DOE-funded Carbon Dioxide Information and Analysis Center (CDIAC), and the DOE-funded Atmospheric Radiation Measurement (ARM) Archive. The EDSS Group also includes the consortium for the development and maintenance of Mercury, which is a state-of-the art toolset for management, harvesting, indexing, and searching of structured spatiotemporal metadata.

Beyond the formal participation outlined in the DOE authorization, the EDSS Group will collaborate with DataNetONE in the following ways:

- Make the Mercury software stack available to DataNetONE and its member nodes for use as a component in the DataNetONE software stack.
- Collaborate with DataNetONE on the further development of Mercury, so that the toolset meets the evolving needs of DataNetONE and the other participants in the Mercury consortium.
- Ensure that the metadata for the holdings of these data centers conforms to the standards established by DataNetONE and that these data center holdings are accessible through the interfaces developed by DataNetONE.
- Test and contribute to the development of the DataNetONE software stack with goal of these data centers becoming full DataNetONE member nodes as soon as that stack reaches an appropriate level of maturity.
- Utilize our considerable experience and expertise in the management of biological, geophysical, and environmental data to help ensure the success and sustainability of DataNetONE.

Sincerely,

Bruce E. Wilson, Ph.D.
Group Leader, Environmental Data Science and Systems
Manager, ORNL NASA Distributed Active Archive Center

March 17, 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

Thank you for the opportunity to collaborate on DataNetONE. The South African National Parks (SANParks) and the South African Environmental Observatory Network (SAEON) both aspire to advance understanding of ecosystems in order to detect, predict, and manage environmental change. We are keenly aware of the need to preserve and share scientific data across earth science disciplines for use in conservation and management of our natural systems, and so we fully endorse the goals of DataNetONE. The network of individuals and institutions that you have assembled for DataNetONE is eminently qualified to create the technical infrastructure for this global network as well as create the sociological setting necessary for widespread participation.

SANParks has been involved in deploying metadata and data management infrastructure across its parks in collaboration with the National Center for Ecological Analysis and Synthesis. We would welcome extensions of this architecture that allowed us to participate in DataNetONE, and we would commit to collaborating with DataNetONE as a Member Node that provides data archival and collaboration services for ecological and environmental research in South Africa. Such a network of Member Nodes that agree upon data, metadata, and communication standards would significantly improve accessibility and interoperability of data resources across the globe and provide massive benefits to environmental science. We look forward to contributing to such an important endeavor.

Sincerely,
Judith Kruger

Program Integrator : Science Support SANParks



addo elephant
agulhas
augrabies falls
bontebok
golden gate highlands
karoo
kgalagadi transfrontier
knysna lake area
kruger
mapungubwe
marakele
mountain zebra
namaqua
table mountain
tankwa-karoo
tsitsikamma
|ai-|ais/richtersveld
vaalbos
west coast
wilderness



行政院農業委員會 林業試驗所
TAIWAN FORESTRY RESEARCH INSTITUTE

臺北市10066南海路53號 電話：02-23039978 傳真：02-23142234 53 Nanhai Road, Taipei 10066, Taiwan Tel/2303-9978 Fax/23142234

March 11, 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The Taiwan Forestry Research Institute (TFRI) and the Taiwan Ecological Research Network (TERN) collect and archive a diverse suite of ecological and environmental data in Taiwan. TFRI and TERN are committed to collaborating with DataNetONE as a Member Node that provides data archival and collaboration services for ecological and environmental research in Taiwan.

There is tremendous value in the development of a shared, distributed network for data from the earth sciences. The DataNetONE goal of providing the distributed framework, sound management, and robust technologies that enable long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data is critical to the goals of long-term ecological research at TFRI and TERN. Our existing systems incorporate data management infrastructure such as Metacat and Morpho that were developed at NCEAS as part of the KNB Network. We will collaborate with DataNetONE to make the extensions of these software technologies robust for use in international, multi-disciplinary systems, particularly addressing issues such as traditional Chinese language support that are important for scientific collaboration.

The group you have assembled for DataNetONE is extremely qualified to build this critical distributed system. We anticipate a strong collaboration between

your team and our science and informatics colleagues. The long-term sustainability of DataNetONE will be key to addressing the large and growing suite of global environmental issues that require synthesis of inter-continental data.

Sincerely,

A handwritten signature in black ink, appearing to read 'Chau Chin Lin' in a cursive style.

Chau Chin Lin

Senior Scientist / Coordinator of Information Management

Taiwan Forestry Research Institute /Taiwan Ecological Research Network

53 Nan Hai Rd. Taipei, Taiwan 100

Ph: +886-2-23039978 ext 2660

Fax:+886-2-23331582



United States Department of the Interior

March 11, 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The United States Geological Survey (USGS) is committed to the long-term management, delivery, and analysis of research data on a national and global scale. The DataNetONE goal of providing the distributed framework, sound management, and robust technologies that enable long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data is very complementary to the overall mission and objectives of the USGS. The proposal also satisfies an immediate need that the USGS National Biological Information Infrastructure (NBII) Program has with respect to meeting its overall objective of serving as an electronic gateway to biological data and information, both textual and spatial in nature, maintained by federal, state, and local government agencies; non-government institutions; and private sector organization in the United States and around the world.

This strong commitment by the USGS is demonstrated by our significant involvement in the initial design, development, and establishment of a long-term sustainable vision for DataNetONE. This is demonstrated through USGS willingness to allocate USGS Program resources, both people and equipment, to the establishment of DataNetONE. USGS contributions in support of the establishment, implementation, and long-term sustainability of DataNetONE are estimated to \$135,000 per year through dedication of staff time to DataNetONE management, leading DataNetONE Working Groups, serving as the short-term Director for Outreach and Education, participating in DataNetONE Executive Board, and the dedication of hardware/software resources for the initial DataNetONE Member Node. USGS will also lead the implementation of DataNetONE standards, protocols, technologies, and archival methods by several federal, state, and local libraries. USGS will accomplish these activities through its federal interagency leadership position in efforts such as CENDI (interagency scientific and technical information organizations), the BioEco Work Group within the Committee on Environment and Natural Resources (CENR), Science.Gov, and through active engagement of over 20 research libraries that exist in USGS. USGS will involve State

data management needs through its leadership in the Organization of Fish & Wildlife Data Managers Association.

USGS also brings several international collaborators, primarily the World Data Center for Biodiversity and Ecology (WDC) and the Inter-American Biodiversity Information Network (IABIN) to the DataNetONE effort. These activates will actively participate in DataNetONE Working Groups, serve as test-beds for DataNetONE implementation actives, and eventually become Member Nodes in the DataNetONE network. USGS will once again represent and provide the resources which allow these global networks to actively participate in all relevant DataNetONE activities.

We look forward to a strong collaboration between your multi-sector team and our colleagues in a number of USGS science and informatics related Programs. The team you have established is very capable of making significant contributions to the advancement of science in this domain and ultimately producing an operational capability that would be of significant value in the management of federal lands, eradication of such issues as invasive species, and in assisting researchers to perform analysis of data that once was only dreamed would be possible. The long-term sustainability of such a network as DataNetONE is key in addressing the pressing science issues that continue to face our Nation. USGS intends to face these challenges with you throughout the life of the project and long into the next century.

A handwritten signature in dark ink, reading "Gladys A. Cotter". The signature is fluid and cursive, with the first name "Gladys" being more prominent than the last name "Cotter".

Gladys A. Cotter

U.S. Geological Survey, Associate Chief Biologist for Information

UNIVERSITY OF CALIFORNIA

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

OFFICE OF THE PRESIDENT
California Digital Library

415 20th Street, 4th Floor
Oakland, California 94612
March 7, 2008

Dr. William K. Michener
Long-Term Ecological Research Network Office, Associate Director
Department of Biology
MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dr. Michener:

The California Digital Library (CDL) at the University of California (UC) is pleased to participate as a sub-awardee in the University of New Mexico's DataNetONE (Observation Network for Earth) project. Digital preservation is one of CDL's cornerstone programs -- UC is committed to maintaining world-class collections of scholarly information. The CDL, in partnership with the UC campuses, established the Digital Preservation Program to ensure long-term access to the digital information that supports and results from research, teaching, and learning at UC. The goals outlined in DataNetONE are an important step in CDL's Digital Preservation Program.

As a project collaborator:

- CDL will contribute our unique perspective as an academic digital library that works collaboratively with the 10 campuses of the University of California to preserve scholarly output
- Patricia Cruse will actively participate in the DataNetONE working Group that is focused on Long-term Sustainability and Governance
- as a co-chair of the Data preservation, metadata, and interoperability working group, John Kunze will help lead this important set of activities
- CDL will provide equipment, space, and systems support for a DataNetONE member node during the project's first year
- CDL will serve as a test site for the library software stack
- CDL will provide knowledge and expertise in the development of online and face-to-face outreach and instructional materials for scientists, students, and citizen scientists
- CDL will provide input and assist in conducting online and local face-to-face instructional sessions for scientists, students, and librarians
- CDL will contribute knowledge and expertise in the development and deployment of a shared, distributed DataNetONE virtual reference service

We look forward to working with the DataNetONE community on this important research endeavor.

Laine Farley
Interim Executive Director
California Digital Library



Bertram Ludäscher
Associate Professor
Dept. of Computer Science & Genome Center
University of California, Davis
One Shields Ave, Davis, CA 95616

Dr. William Michener
Long Term Ecological Research Network Office
University of New Mexico
MSC03 2020
Albuquerque, NM 87131-0001

Re. Support for DataNetONE

March 17, 2008

Dear Bill,

It is my pleasure to provide this letter of collaboration for your DataNetOne project. As the director of the Data and Knowledge Systems (DAKS) Lab at the UC Davis Genome Center and the Department of Computer Science, I would like to confirm my intent and enthusiasm to collaborate with you and the DataNetOne participants.

The DAKS lab is conducting research and development on scientific data and workflow management and knowledge-based extensions to facilitate data integration. For example, in the *Kepler/CORE* project we develop a comprehensive, open, robust, and extensible scientific workflow infrastructure based on the Kepler scientific workflow system; in *Kepler/ChIP-chip*, we are developing a collaborative scientific workflow environment for accelerating genome-scale biological research; in *Kepler/pPPOD* we are developing a provenance-aware, collection-oriented system for the ATOL community; and in *ITR/SEEK* we have developed semantic mediation technology to help scientists integrated ecological data. My lab currently consists of two project scientists, two developers, two postdocs (being recruited), and several graduate and undergraduate students.

I am excited to co-lead the *Data Integration and Semantics* Working Group and to participate in the *Scientific Workflows* Working Group. While at the San Diego Supercomputer Center (from 1998—2004) I have also worked on digital library projects, and developed an approach for knowledge-based archival of data collections that takes into account semantic techniques, workflow approaches, and the Open Archives Initiatives Standard. DataNetOne brings together leading experts in all relevant areas of DataNet and I am looking forward to be a part of it.

Sincerely,

A handwritten signature in cursive script that reads 'Bertram Ludäscher'.

Bertram Ludäscher, Associate Professor
Department of Computer Science & Genome Center
University of California at Davis
ludaesch@ucdavis.edu

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

BERKELEY • DAVIS • IRVINE • LOS ANGELES • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

PHONE: (805) 892-2500
FAX: (805) 892-2510

SANTA BARBARA, CALIFORNIA 93106-6150

March 17, 2008

Dr. William Michener
Department of Biology, 1 UNM
MSC03 2020
Albuquerque, NM 87131-0001 USA

Dear Dr. Michener,

At the University of California Santa Barbara, we believe that the DataNetONE initiative provides a critical opportunity to link the environmental data repositories being developed in the US and globally at a time when the need for integrated global data systems for the environment has never been greater. The University of California Santa Barbara has been a leader in data management solutions for the earth and environmental sciences. Programs at UCSB such as the National Center for Ecological Analysis and Synthesis, the Ecoinformatics Center at the Marine Science Institute, the Map and Imagery Laboratory at Davidson Library, and the Alexandria Digital Library Project have pioneered software systems for archiving and preserving scientific data. The Knowledge Network for Biocomplexity (KNB) is a data archive hosted by NCEAS that holds over 15,000 environmental data sets, and the MIL and Alexandria Project house metadata records for millions of maps and geospatial images. These archiving efforts demonstrate UCSB's long-term commitment to scientific data management, and are part of UCSB's informatics initiatives to develop standardized approaches for managing data and metadata and analysis and modeling tools that are useful for science and policy development.

The goals of the DataNetONE initiative, such as the goal of providing the distributed framework, sound management, and robust technologies that enable long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data, are very consistent with the mission and objectives of UCSB's various informatics programs. For these reasons, we believe that DataNetONE will make fundamentally important advances in cyberinfrastructure for ecology and other earth sciences. UCSB is committed to collaborating with DataNetONE through numerous activities. First, we have made significant investments in the initial design, development, and establishment of a long-term sustainable vision for DataNetONE through the involvement of Matthew Jones (Director of Informatics Research and Development) and Stephanie Hampton (Deputy Director) at NCEAS. Second, UCSB will house one of the three national Coordinating Nodes by providing:

- Sufficient space, power, network, and air conditioning for the DataNetONE computing and networking equipment in the machine room at the Davidson Library, as described in the Cyberinfrastructure appendix of the proposal. The Davidson Library is a strong partner with the California Digital Library, and this role in preserving scientific data for future generations of researchers strengthens our library preservation programs.
- Office space and facilities in the Marine Science Building for DataNetONE personnel, including software engineers and systems administrators that will be designing, developing, and deploying infrastructure for DataNetONE
- Office space and facilities at the National Center for Ecological Analysis and Synthesis (NCEAS) for postdoctoral researchers and students who are contributing to the research and education activities of DataNetONE.

Finally, UCSB personnel will play key leadership roles in DataNetONE. Stephanie Hampton and Matthew Jones will serve on the initial External Advisory Committee, and Dr. Hampton will co-lead the Community engagement and education working group. Co-PI Matthew Jones will serve part-time as co-leader of the Core Cyberinfrastructure Team (CCiT) and will share responsibilities of the Assistant Director for Development and Operations on an interim basis until that position is filled. Co-PI Stephanie Hampton will similarly share responsibilities of the Assistant Director for Community Engagement and Outreach part-time until that position is filled.

NCEAS and other programs at UCSB support synthetic and collaborative research in ecology and environmental biology by scientists throughout the US and the world. Due to the extensive re-use of existing data for these synthesis activities, we have recognized the critical nature of ecological informatics to advances in ecology and environmental biology. We believe that linking environmental data networks into a broad coalition of data centers via DataNetONE would be a tremendous leap forward for environmental science in the US and globally. The standards, technologies, and infrastructure being developed by you and your partners will be key to delivering and improving overall access to global environmental data holdings. I encourage your continued efforts in this area and look forward to a fruitful and mutually beneficial collaboration on DataNetONE.

Sincerely,



William Murdoch
Interim Director, NCEAS



Mark Brzezinski
Director, Marine Science Institute



Brenda Johnson
University Librarian

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

National Center for Supercomputing Applications

1008 NCSA Building
1205 West Clark Street
Urbana, IL 61801



March 19, 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign has, for over 20 years, been a National Science Foundation (NSF) supercomputing center. Looking forward, NCSA is also the site of the upcoming NSF Track 1 system which will deploy a sustained petascale computational resource and associated data archive for the national research community. NCSA has, and will for the foreseeable future, continue to have some of the most advanced computational, storage, and networking resources in the nation.

NCSA has long had an interest in and involvement with the development cyberinfrastructure for the ecological and environmental research communities. Our understanding of the DataNetONE program is that, first, it should be scalable to the largest scales in most, if not all, discipline areas. Second, there is a need to put DataNetONE on a sustainable basis. For DataNetONE this will take the form of creating structures and models to provide data services for collections and users, both large and small, and to simultaneously secure funding streams to support this infrastructure.

We are pleased to collaborate with you and your team's DataNetONE proposal. In particular NCSA is committed to act in two planned roles:

- Randy Butler and Von Welch will lead the Federated Security working group for DataNetONE.
- Once DataNetONE secures infrastructure funding streams, NCSA will engage in a process to act as a large data store for DataNetONE on a sustainable basis.

In addition, as a member of the DataNet team, NCSA will also collaborate, where appropriate, with other members of the team in areas of mutual interest.

We look forward to your success.

Sincerely,

Director, National Center for Supercomputing Applications
Professor and Distinguished Chair for Research Excellence, Department of Chemistry
University of Illinois at Urbana-Champaign

xc: Danny Powell, Executive Director, National Center for Supercomputing Applications

UNIVERSITY OF ILLINOIS
AT CHICAGO

University Library (MC 234)
Box 8198, Chicago, Illinois 60680

March 5, 2008

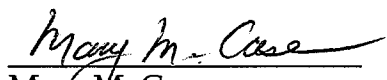
Dr. William K. Michener
Long-Term Ecological Research Network Office, Associate Director
Department of Biology
MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001 USA

Dr. Michener:

The Library at the University of Illinois at Chicago (UIC) agrees to collaborate as a sub-awardee in the University of New Mexico's DataNetONE (Observation Network for Earth) project. As a project collaborator, UIC will provide

- equipment, space, and systems support for a DataNetONE member node during the project's first year
- serve as a test site for the library software stack
- assist in the development of online and face-to-face outreach and instructional materials for scientists, students, and citizen scientists
- conduct online and local face-to-face instructional sessions for scientists, students, and librarians
- assist in the development and operation of a shared, distributed DataNetONE virtual reference service
- provide server, technical, and editorial resources for an open-access online journal to support reporting of research and operational findings from DataNetONE and other DataNet partner projects
- Dr. Robert J. Sandusky will serve as a member of the Core Cyberinfrastructure Team

We look forward to a rewarding and productive research effort.


Mary M. Case
University Librarian
Richard J. Daley Library
University of Illinois at Chicago

UIC



March 10, 2008

Dr. William K. Michener, Associate Director
Long-Term Ecological Research Network Office,
Department of Biology, MSC03 2020
The University of New Mexico
Albuquerque, NM 87131-0001 USA

Dear Dr Michener,

It is a pleasure to commit the resources of the Biodiversity Research Center (BRC) at the University of Kansas as a collaborator in the DataNetONE consortium. The BRC holds voucher collections of more than 8 million specimens of animals and plants that document the life of the planet, past and present, from genomic information to biogeographic occurrences of individual organisms, populations and species, to their morphological, behavioral and ecological characteristics. Much of these data has been captured and archived digitally and shared in community biodiversity networks.

For more than a decade, the BRC has designed, implemented and promoted software solutions to enable the digital preservation (the NSF-funded *Specify* project) and network dissemination of those data in community-defined standard formats of the Internet through various protocols developed at the BRC and elsewhere.

The BRC welcomes the opportunity to be a part of the DataNetONE consortium to catalyze the long-term preservation of these irreplaceable data, and particularly, advance the semantic interoperability of information collected, housed and archived by biodiversity research facilities such as natural history museums. The research and infrastructural goals of the DataNetONE initiative are well-aligned with the goals of the BRC. As such,

- Dr. David A. Vieglais, Senior Scientist at the BRC, will serve as a member of the Core Cyberinfrastructure Team
- the BRC will serve as a test site for a DataNetONE Member Node
- the BRC will assist in the development of service interfaces between biodiversity collection management software such as *Specify* and the DataNetONE archive software stack
- As a Board Member of the National Science Collection Alliance (NSCA), I will be a liaison between the NSCA and the DataNetONE project to facilitate the collaboration of NSCA's more than 100 biocollection facilities collectively holding more than 500 million specimens of animals and plants and their associated biodiversity and ecological data.

We look forward to a rewarding and productive research partnership.

Best,

A handwritten signature in cursive script that reads 'L. Krishtalka'.

Leonard Krishtalka
Director
krishtalka@ku.edu



The Department of Computer Science
The University of Manchester
Kilburn Building
Oxford Road
Manchester M13 9PL
UK

Tel: +44 161 275 6195
Fax: +44 161 275 6236
Email: carole@cs.man.ac.uk
12th March 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Bill

Re: DataNetONE

I am delighted to offer you the support of The University of Manchester for the DataNetONE project proposal. We enthusiastically support the aims of DataNetONE and we welcome this opportunity to participate in this bid.

Manchester is a key national e-Science centre and was one of the original e-Science Centres set up in 2002 to found the backbone around which a permanent UK e-Infrastructure could grow. It is unique in being a Computer Science (research-oriented) and Manchester Computing (service-oriented) venture from the start. We now also host the National Centre for e-Social Science (NCeSS) and the National Centre for Text Mining (NaCTeM). We are a founding member of the UK's National Grid Service.

The Open Middleware Infrastructure Institute UK (OMII UK) node in Manchester provides a team of software engineers and support staff to develop the highly popular open source Taverna Workflow Workbench and Workflow Engine. Taverna was designed for linking together resources in the Life Science and BioMedical community, and is used by over 300 organisations internationally for routine bioinformatics and biomedicine including systems biology. In conjunction with University of Southampton we have developed the myExperiment social networking environment for sharing and reusing the digital artifacts of science, including scientific workflows.

We have very extensive experience in the data arena. The Information Management Group conducts research into the design, development and use of data and knowledge management systems. Such research activities are broad in nature as well as scope, including basic research on models and languages that underpins activities on algorithms, technologies and architectures. Challenging applications motivate and validate our research, in particular the Semantic Web and e-Science.

With our significant role in national e-Science infrastructure, expertise in information management and extensive experience of working with an international user base, we bring skills which complement the DataNetONE project. We look forward to working with such a strong consortium of partnering institutions in this important and exciting project.

Yours sincerely

A handwritten signature in black ink, appearing to read 'Carole Goble', written in a cursive style.

Professor Carole Goble
The University of Manchester, UK
Director, ^{my}Grid Project and ESNW



The University of New Mexico

Office of the Vice President for
Research and Economic Development
Scholes Hall Room 327, MSC05 3480
1 University of New Mexico
Albuquerque, NM 87131-0001
Telephone (505) 277-6128
FAX (505) 277-5271

5 March 2008

Dr. William Michener
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
Department of Biology, MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Bill:

The University of New Mexico is pleased to serve as the host institution and one of the three Coordinating Nodes for the DataNetONE proposal that you and your distinguished colleagues are submitting to the National Science Foundation. The University has a long research tradition in biology, ecology, and the environmental and earth sciences. Moreover, the State of New Mexico and UNM are making a significant investment in cyberinfrastructure via the new world-class supercomputer that will be housed at Intel Corporation and the similarly-architected exemplar HPC machine that is being installed on campus.

There are tremendous synergies that can be realized by locating the DataNetONE central office at UNM and the Coordinating Node in association with the New Mexico Computing Application Center and Intel Corporation. Consequently, UNM offers the following significant financial support to DataNetONE:

- Office space at the UNM Science and Technology Park. This space includes 7 large offices, several of which can be partitioned to support two or more staff members or students, as well as a spacious conference room, and receptionist/open space. This commitment includes support for utilities and internet access, plus access to two additional nearby conference rooms for two weeks each year of the five-year project. It should also be noted that the office complex is presently occupied by the Joint Technology Office (JTO) and offers a high level of security and access to ample parking. JTO is slated to move into their new building on Kirtland Air Force Base on or about September 2008; interim space is available at the S&T Park should their move be delayed for any reason.
- A budget of \$120,000 to support office furnishings.
- Six weeks salary support for you as PI for each year of the project to participate in DataNetONE Working Groups and related activities (a contribution of \$128,901 over the lifetime of the award which includes salary and fringe benefits).

Please keep me apprised of the status of DataNetONE. I look forward to working with you and your team to make this important national facility a reality.

Sincerely,

John K. McIver
Interim Vice President for Research and Economic Development

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

12th March, 2008

Dear Bill

DataNetONE Proposal

The University of Southampton plays a significant role in the UK e-Science programme, with a distinctive emphasis on the data lifecycle, data reuse and data curation. We are very pleased to collaborate with DataNetONE, with its goal of providing the distributed framework for long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data.

From the inception of e-Science, when Southampton participated in three of the UK's six Pilot Projects, we have worked across multiple disciplines in the recording and reuse of scientific data, focusing on the lifecycle from sensor network and laboratory bench through to publication and repositories. With 18 multidisciplinary e-Science projects led from Computer Science and working directly with application domains, and a further 17 led directly by those domains, we have very extensive experience in the handling of scientific data as well as in supporting researchers in its use.

In all our work in e-Science and open repositories we make extensive use of Web technologies, including Semantic Web for linked data and provenance and Web 2.0 for rapid application development and community curation. In collaboration with The University of Manchester we pioneered the Semantic Grid initiative, which has had significant influence in bringing together Web and Grid technologies in Europe and internationally. The Web Science Research Initiative (WSRI) is a joint endeavour between the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT and the School of Electronics and Computer Science (ECS) at the University of Southampton.

Nationally, Southampton led the six year Advanced Knowledge Technologies programme, an interdisciplinary research collaboration to realise the opportunity of our information-rich environment, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and involving internationally recognised research groups at the Universities of Aberdeen, Edinburgh, Sheffield and the Open University. We collaborate in the South East region of England with the Oxford e-Research Centre, the Reading e-Science Centre (which is supported by the UK Natural Environment Research Council) and the e-Science Centre of the UK Science and Technology Facilities Council.

We are excited to participate in DataNetONE, which we believe is set to make the significant contribution to the advancement of science which is only possible through taking this data-centric approach and such a strong group of collaborators. We look forward to bringing the experience and practice established through 8 years of e-Science to benefit DataNetONE and the researchers that use it.

Yours sincerely

A handwritten signature in blue ink, appearing to read 'David De Roure', with a stylized, cursive script.

David De Roure
Professor of Computer Science

Direct tel: +44 (0)23 8059 2418
email: dder@ecs.soton.ac.uk



March 7, 2008

Dr. William Michener
Long Term Ecological Research Network Office
University of New Mexico
MSC03 2020
Albuquerque, NM 87131-0001

Dear Bill,

It is my pleasure to provide this letter of collaboration for your DataNetOne project. On behalf of the Pegasus Workflow Management System project, I would like to confirm my intent and enthusiasm to collaborate with you and the DataNetOne participants on issues related to scientific workflow management.

Pegasus-WMS is a flexible framework that maps abstract application workflow descriptions to executable workflows that can be executed in a distributed environment such as the Grid. The scientific applications represented by the workflows are often large-scale and complex and may contain hundreds of thousands of tasks. Pegasus may map the entire workflow at once or portions of it. The workflow execution engine of Pegasus-WMS (Condor DAGMan) provides scalability and fault tolerance when executing large-scale, data-intensive workflows in distributed environments. Pegasus-WMS is used to manage applications from scientific domains such as astronomy, biology, earthquake science, and gravitational-wave physics, and others.

As part of DataNetOne, I am committed to participate in the project activities as the co-leader of the Scientific Workflows working group. Today, there are many scientific workflow systems, including Kepler, Taverna, Pegasus, and others. However, each has its own representation and its own way to capture and store provenance information. As we look into the future and deal with issues of data preservation, it is critical to start defining interoperable workflow representations which can be used to archive workflow-based analyses descriptions. In order to be able to interpret data in the long term, interoperable provenance formats are needed. I believe that working with the co-leads of the Scientific Workflow group (and leaders of the Taverna project—Carole Goble and David DeRoure) as well as with the many members of DataNet working on the Kepler system and related scientific workflow issues, we can accomplish longer-term, sustained interoperability in a number of workflow system facets. We can also bring our

collective experience with applications in a variety of scientific domains to articulate workflow design patterns prevalent in the scientific workflow applications.

In summary, the DataNetOne plans for research into long-term preservation of diverse and complex multi-scale, multi-discipline, and multi-national science data and into scientific workflows are very exciting and I hope to continue our collaboration in the various facets of the problem.



Ewa Deelman, Ph.D.
Research Assistant Professor of Computer Science, University of Southern California
Project Leader, USC/Information Sciences Institute

Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 (USA)
Phone: +1-310-448-8408
Fax: +1-310-822-7719
<http://www.isi.edu/~deelman>



March 10, 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The University of Tennessee, College of Communication and Information's School of Information Sciences (UT) is pleased to collaborate with you and to serve as a Member Node in the DataNetONE project. As a member node, UT is committed to providing library and information science support to the DataNetONE project and to coordinating participation of the UT, University of Illinois-Chicago, and other libraries.

UT's commitment to DataNetONE has already been substantial through our personnel's participation in the initial design of DataNetONE including technological, sociological and financial aspects of the organization. The School of Information Sciences is committed to engagement in both technological and sociological aspects of this project and will be working with Oak Ridge National Lab, a DataNetONE Coordinating Node, supporting hardware installation and software development. SIS will be participating in organizational development and serving as a project center for research focusing on library and information science solutions to long term digital preservation of earth observational data. Faculty will provide technological and research expertise. Dr. Bruce Wilson will be significantly involved in initial cyber infrastructure design and development at the start of the project, then will continue as a member of the Core Cyberinfrastructure Team. Dr. Suzie Allard will coordinate research based at the School and will also serve as a co-leader on the SocioCultural Working Group. Other faculty members will also serve on key committees in the DataNetONE organization. Dr. Carol Tenopir is a co-leader of the SocioCultural Working Group, and Dr. Lorraine Normore is a co-leader of the Usability and Assessment Working Group.

In addition, UT will draw on a rich pool of information sciences graduate student for graduate research assistants and alumni and will coordinate hiring of students, programmers and a post-doctoral assistant.

UT is committed to collaboration and providing leadership in the library and information sciences to many aspects of the DataNetONE project.

Sincerely,

Michael O. Wirth, Dean



National Coordinating Office
1955 East 6th Street
Tucson, AZ 85721
Phone: (520) 626-3821
Fax: (520) 621-3816
www.usanpn.org

6 March 2008

William K. Michener, PhD
Director, New Mexico EPSCoR State Program
Associate Director, LTER Network Office
NM EPSCoR, MSC05 3180
1717 Roma, NE
1 University of New Mexico
Albuquerque, NM 87131-0001

Dear Dr. Michener:

The USA-National Phenology Network (NPN) is an emerging and exciting partnership between federal agencies, the academic community, and the general public to monitor and understand the influence of phenology on the nation's resources. The goal of the NPN is to establish a continental-scale science and monitoring program focused on phenology – the periodicity of plant and animal life cycles driven by seasonal variations in climate. The NPN will capitalize on integration with other monitoring efforts, remote sensing platforms and products, emerging technologies and data management capabilities, formal and informal educational opportunities, and a new readiness of the public to participate in investigations of natural systems on a national scale.

The DataNetONE goal of enabling long-term management and preservation of diverse cross-scale and multi-discipline observational data obviously complements the overall mission and objectives of the NPN. In addition, the DataNetONE program would help us meet our goal of serving as a clearinghouse for biological data on plant and animal phenology, maintained at the NPN as well as in numerous governmental and non-governmental agencies and organizations in the United States.

One important aspect of the NPN is our citizen science, education and outreach program. Our goal is to create a range of programs and products designed to (1) engage the public in long-term phenological data collection and analysis through formal and informal science education programs, (2) engender self-directed, voluntary learning using inquiry-based approaches, (3) provide training in the tools and applications of phenological studies to citizens and scientists, and (4) enhance opportunities for the public to interact with professional scientists. We will attempt to reach a diverse audience through a

variety of programs that encompass multiple levels of participation and engagement. It is our goal that data collected by citizen scientists be an integral part of the NPN database; as such, it is in our interest to ensure that data quality and standards are very high, and that data are managed and archived properly.

As such, I am pleased to participate in the Citizen science and public outreach working group of DataNetONE, where we will establish requirements for management of citizen science data and visualization, exploration, and analysis of data by a variety of users, create a comprehensive data management strategy for highly disparate citizen-based observational networks, and build tools to allow project managers, researchers, educators, or networks to develop a customizable web-based data gathering system. These activities are closely aligned with the critical citizen science component of our network, as well as our NSF Research Coordination Grant goal of enabling researchers and interested citizens to store, discover, and retrieve phenological data from a distributed network of databases.

In sum, the DataNetONE network will be key to address critical ongoing and potential future science and environmental issues that face the US, and a collaboration between DataNetONE and the NPN should be mutually beneficial, and would ultimately have important impact on US science programs and policies related to global change. We look forward to a strong collaboration between your team and the staff here at the NPN on this exciting venture.

Sincerely,

A handwritten signature in dark ink, appearing to read "Jake Weltzin", with a stylized, cursive script.

Jake Weltzin
Executive Director
USA National Phenology Network



SPONSORED PROGRAMS OFFICE

1415 Old Main Hill
Old Main Room 64
Logan, UT 84322-1415

March 6, 2008

Dr. William K. Michener
Long-Term Ecological Research Network Office, Associate Director
University of New Mexico
Department of Biology
MSC03 2020
Albuquerque, NM 87131

Subject: Statement of Intent to Collaborate "DataNetONE (Observation Network for Earth)"
USU Proposal Code: 080768

Dr. Michener:

Utah State University (USU) is pleased to be a collaborator with the University of New Mexico (UNM) on the NSF proposal. USU agrees to provide those personnel and resources identified by our PI, Jeffrey Horsburgh, in support of this effort. We understand that UNM, will be the prime contractor and USU will be a subcontractor in the event of an award.

The mission of the USU's Utah Water Research Laboratory (UWRL) includes conducting research that is directed at solving multimedia water-related problems of state, national, and international scopes. We certainly believe that the goals of the DataNetONE project and the long term disposition of water resources and environmental engineering related data fit within this mission and are also consistent with our goal to generate, transmit, apply, and preserve knowledge.

USU brings a strong linkage to the hydrologic and environmental engineering research communities through the involvement of Jeffery Horsburgh. His involvement in the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System project as well as his involvement as a Co-PI on one of the Water and Environmental Research Systems (WATERS) Environmental Observatory Test Bed projects uniquely situates him to participate as a member of the Core Cyberinfrastructure Team. As a DataNetONE project collaborator, USU will provide the following through the involvement of co-investigator Jeffery Horsburgh:

- Service as a member of the Core Cyberinfrastructure Team
- Attendance at the meetings of the Core Cyberinfrastructure Team
- Intellectual contribution to the design and implementation of DataNetONE architecture
- Service as a liaison between DataNetONE and Utah State University as well as other relevant organizations



We appreciate your interest in using the resources of USU and look forward to working together on this effort. Please feel free to direct questions of a technical nature or regarding the Statement of Work to Jeffrey Horsburgh at (435) 797-2946. Questions of a contractual or administrative nature should be directed to the undersigned at (435) 797-1659 or norma.buxton@usu.edu.

Sincerely,

A handwritten signature in blue ink, appearing to read 'NB' followed by a stylized flourish.

Norma Buxton
Sponsored Programs Administrator

cc: File